Understanding Booksonomies:

How and why are book taggers tagging?

Aedín Guyot

A dissertation submitted to the University of Wales in partial fulfilment of the requirements for the degree of Magister in Scientia Economica (MSc) under Alternative Regulations

Department of Information Studies

University of Wales

Aberystwyth

2013

Abstract

Tagging is an ever-growing feature of online systems. As more and more content is tagged by users, the resulting "folksonomies" grow and can become unwieldy. The design of tagging systems must take into account how the resulting networks of tags are composed, and what motivated the taggers, in order to best use those tags as an aid toward search on the system.

A literature review was carried out on the topics of folksonomies in general, how they compare with more formal ontologies and how folksonomies can be improved. Studies categorising folksonomy tags were analysed, with particular attention paid to those studies using the resulting categorisation information as a means to infer tagger motivation. A specific strand of the literature review focussed on studies of book-tagging systems.

The aim of this study was to take a particular tagging system, the book website LibraryThing, and analyse the tags on fifty sample books. Long tail tags with a frequency of 2 or less, were ignored for reasons detailed in the research methodology, leaving a total of 13,358 tags to be viewed and categorised. The tag frequency distribution was shown to demonstrate the Zipfian power law. The tag categorisation model indicated that booksonomy taggers generally tag within the categories of "genre/style", "subject" and "personal task-based". Users are motivated mainly by their own personal organisational needs, but also by some social impulses towards the other users of the site. As a further component of the study, tags on two specific book genres (non-fiction and young adult) were analysed separately. Patterns such as higher-than-average tagging in the "target reader" category on young adult books became apparent.

The original research and datasets generated for this study provide further source material as a contribution to the evolving discussion on folksonomies in general, and on booksonomies in particular.

Declaration and Statement of Originality

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This work is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes and/or an appropriate citation method giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give consent for my work, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Table of Contents

List of Figures/Tables/Charts

Acknowledgements

I would like to thank my supervisor Pauline Rafferty for her support during the dissertation process. Other Aberystwyth University staff members who were helpful in answering my various questions were Sue Lithgow and Juanita Foster-Jones**.**

Many thanks go also to my mum Bríd, brother and sisters Brían, Denise and Maeve for their kindness and encouragement, always.

And finally, I couldn't have done it without the patient and loving support of my husband Olivier and the welcome distraction of my toddler son Max, not to mention baby-to-be whose due date falls one month after this dissertation's submission date!

1 Introduction

1.1 Overview

Social tagging capabilities are now very common on networked systems and websites, allowing users to add words and phrases to content, and to search and browse the words and phrases added by other users. Folksonomies, another word for these systems of tags, are considered by many to be a crucial part of today's information landscape and indispensable in providing a fulfilling user search experience. According to studies carried out on tagging behaviour, there are generally two broad motivations for an individual to tag content – the first is for personal recall and organisation purposes, and the second is with a more social purpose in mind, to provide information about the item to other people, and thus to connect on some level with those people. There are numerous systems on the internet that allow tagging of various types of items, among them products, films, people, images and books. With many users tagging many items, a system of tags becomes built up over time, forming what has become known as a "folksonomy".

A number of websites have developed over tagging's lifetime specifically allowing for the tagging of books. These sites include Shelfari, LibraryThing, GoodReads and aNobil. This study aims to analyse the tags that people apply to books, to build a category model to demonstrate the patterns within those tags, to demonstrate the frequency distributions for the applied tags and categories, and to build on this analysis to try to assess what taggers' motivations might be in applying these tags. In essence, this study attempts to ascertain what a "booksonomy" contains, and to infer to some degree why book taggers tag.

The LibraryThing website is "an online service to help people catalog their books easily" (LibraryThing, 2012). Although the main stated purpose of the website is to allow the cataloguing of books, the social aspects of the website are also clear, with many groups having formed amongst those members, and with reviews, recommendations and indeed tags themselves forming a rich information source for members and visitors. LibraryThing was chosen for the study partly due to its popularity among users; according to the LibraryThing website (LibraryThing, 2012), there were 1.6 million registered LibraryThing users by January 2013, who between them had added 93.5 million tags to 78.3 million books. As well as its popularity, LibraryThing was also a natural choice due to the availability of its tags in a format conducive to data retrieval and normalisation and the availability of additional

information to aid in the selection of books for the process of analysis and to inform that analysis.

To analyse the data, the tags applied by all users on fifty of the most popular books on LibraryThing were divided into broad categories such as "genre/style", "personal task-based", "character/setting information" and so on. In total, twenty four categories became apparent during the analysis. The frequency of application of each tag was taken into account, ensuring that this "weighting" was reflected in the final category totals. The tag frequency distribution was also assessed to discover if it followed a Zipfian power-law distribution, as proposed in the literature. Finally, a discussion of what the breakdown of categories within the booksonomy might imply about users' motivations for tagging books was undertaken. Understanding such motivations could allow for the better design of tag recommendation systems and tag hierarchies, improving the functionality of book folksonomies in general.

1.2 Aims and Objectives

1.2.1 Aim

To investigate the tags applied to books by website users in order to define a category model for a folksonomy specifically containing book tags (a "booksonomy"), with a view to understanding the motivations behind its users' tagging behaviours.

1.2.2 Objectives

1. To undertake a review of the scholarly literature regarding folksonomies and user tagging decisions and purposes.
2. To analyse a sample set of book tags applied by multiple users and to categorise those tags, thus building up a category model of a "booksonomy". The category model to take into account frequency of application of tags as well as tag counts.
3. To assess whether the tag frequency distribution follows the Zipfian power-law distribution model.
4. To assess how the book tag category model compares with category models suggested by the literature, for books and for other resources.
5. To discuss what the categorisation of the tags might imply about taggers' motivations when tagging books.

1.3 Scope and definitions

1.3.1 Scope

The scope of this research was limited by the resources available for carrying out analysis of tags. A sample set of fifty books was analysed, with the sample set selected based on "most reviews" on the LibraryThing website. This sample set was chosen as it generally is the case that the more reviews a book has, the more tags it also has, and so the sample set provided a substantive set of tags for analysis. However, choosing the books in this way might further bias the selection towards particular types of books, those that tend to be reviewed more than others. Who the typical LibraryThing user is would also be a limiting factor to how representative this research can be considered to be.

Due to only one researcher carrying out the categorisation, the research is prone to a high level of subjectivity. Furthermore, due to resource constraints, it was only possible to categorise tags with an application frequency of greater than 2. This means that understanding the true "long tail" of a "booksonomy" is beyond the scope of this research. Finally, analysis of a larger number of books than fifty would be required in order to arrive at a more statistically accurate breakdown of booksonomies in general. A larger scale research project would be able to in particular improve statistical accuracy for the breakdown of categorisation results by book type, which could only be analysed in this study on the small number of books of each book type available within the initial sample size of fifty.

1.4 Structure

The dissertation is organised as follows – in Chapter 2, an overview of related research and literature is given. Chapter 3 provides details of the methodology followed during the procurement and normalisation of data and the categorisation of that data, with Chapter 4 giving the results of that categorisation. A discussion of what the categorisation results might imply follows in Chapter 5, with Chapter 6 summarising the study's findings in the context of the literature review, and proposing conclusions and possibilities for future work.

1.5 Referencing

Throughout this dissertation, the Harvard American Psychological Association (APA) style of referencing and citation is used.

2 Literature Review

2.1 Introduction

It was Thomas Vander Wal who coined the term "folksonomy", combining the two words "folk" and "taxonomy" to describe the relatively new phenomenon at the time (2004) of users freely adding tags to online content (Vander Wal, 2007). The purpose of this review is to examine existing research into the general theories and definitions within the realm of tagging and folksonomies, into the categories into which user tags fall, and the motivations for tagging that this categorisation might reveal. Studies that focus on media types such as images, films and websites are examined, as are the relatively small number of studies that have been carried out on tag collections for books. In the subsequent chapters, the substantive new data analysis carried out for this research project will be integrated into the literature review presented here.

2.2 Comparing folksonomy-based systems with more formal ontologies

Much research has focussed on comparing folksonomies with more traditional ontology-based information retrieval systems and discussing the advantages and disadvantages of each type of system. Generally, the main advantages of folksonomies are seen to be their flexibility, responsiveness and inclusiveness, with the disadvantages relating to precision and recall in information retrieval due to issues like ambiguity and lack of synonym linking or hierarchical information.

Clay Shirky (Shirky, 2005), one of the "most outspoken proponents" of folksonomies in the literature (Wichowski, 2009) discusses folksonomies versus formal ontologies and the idea that the internet cannot be categorised in the traditional library sense of the word, but must allow users to search and organise organically. He mentions the search engine Google, and discusses the idea that Google was adopted so quickly by previous users of Yahoo and other "search engines" because "Google understood there is no shelf, and there is no file system" in that post-coordination suits user internet search much better than pre-coordination. Shirky also discusses "signal loss", or the loss of information for the user that would occur if the multiple tags that make up folksonomies were to be overly-condensed into a more ontological structure and believes that "the only group that can categorize everything is everybody", so that the individual user's "search question" can change from "Is everyone tagging any given link 'correctly'" to "Is anyone tagging it the way I do?". Smith (2008) in

line with Shirky, also talks about the information loss that would occur if tags were overly synonymised, or bundled into groups "if you treat movies and cinema as synonyms you're ignoring what we might call their sociosemantic differences".

The additional information provided by user tags when compared with more formal indexing techniques is a focus of other studies (Lu, Park, & Hu, 2010; Iyer & Bungo, 2011; Bates & Rowley, 2011; Lawson, 2009; Heckner, Neubauer, & Wolff, 2008). Bates & Rowley (2011), for example, study how different "worldviews" can be accommodated by folksonomies where more formal ontologies may fail, by analysing how the tagging of "LGBTQ" books differs between user tags on book tagging site LibraryThing and expert-assigned tags in library catalogues. They argue that certain types of term work better in folksonomies as they can be applied to resources by a community of users who have a particular awareness of the context of the resource or "collective knowledge domain", which is unlikely to be equalled by "expert" cataloguers with just a broad general knowledge of the context. The LibraryThing tag base is praised as "an organic, deep and dynamic collection of subject metadata in everyday language, created by people that have read the books and who are participants in diverse "lifeworlds" with multiple worldviews".

Not all studies take a fully favourable view of folksonomies, however. The problems that can arise with user tags such as polysemous words, synonymous words, misspellings and personal task management related tags that have little relevance to others are focussed on by several studies (Golder & Huberman, 2006; Peters & Weller, 2008; Bischoff, Firan, Nejdl, & Paiu, 2008) and these and other studies also discuss how folksonomies can be improved and leveraged in order to provide users with a more useful information retrieval experience.

2.3 Improving folksonomies

As mentioned above, various researchers, having discussed the issues with folksonomies, then attempt to find ways to overcome those issues. Guy & Tonkin (2006) look at the issues that "untidy" tags can introduce into tagging systems, in which "tags are often ambiguous, overly personalised and inexact", and means by which improvements can be made in users' tagging behaviours and tagging systems as a whole. These include, for example, suggestions to users while tagging, and the creation of "tag bundles" to bring semantically related tags together for the purposes of improving search recall. Caution is advised, however "There is a real danger that by tidying up tags we are condoning the implementation of a destructive

solution that may lose valuable metadata". Spiteri (2007) reviews sets of tags from three folksonomy sites and analyses the "quality" of tags using the National Information Standards Organization's guidelines for controlled vocabulary construction as a standard. Overall, the tags were found to be quite well-formed, with the main issues being ambiguity, inconsistent use of singular and plural, and unqualified abbreviations or acronyms. Spiteri's conclusion was that some limited education of users who tag would be beneficial in arriving at a folksonomy that could serve search more robustly. The need for intervention by systems or experts in order to make folksonomies more useful is also discussed by Peters & Weller (2008) who use gardening as an analogy in order to discuss maintenance that might be carried out on folksonomies in order to make them more useful. They discuss the benefits and weaknesses of folksonomies, stating that in folksonomies, the "lack of vocabulary control is the price for facile usability, flexibility and representation of active and dynamic language". They also propose treating different types of tags differently within a folksonomy system, taking into account for example whether a tag is a "content" tag or an "organizational" tag and altering its display to the user accordingly. As mentioned by Wichowski (2009), "one of the main problems with tags in folksonomies is the absence of context". In another study within that year Overell, Sigurbjörnsson, & van Zwol (2009) investigate a means of automatically classifying Flickr tags into semantic categories, and thus providing this context, by building a classification system based on Wordnet and Wikipedia articles and mapping tags to this system. A similar approach to providing tag context is outlined by Suchanek (2008).

2.4 The "long tail"

Zipf (1935) demonstrated that when the words in a given corpus are analysed, the plot of word against frequency of the word follows a power law curve. This means that the most frequent words form a huge proportion of the corpus, with the less frequent words tailing quickly off in frequency. In a folksonomy context, this implies that the more popular tags will appear at a far higher frequency than the less popular tags.

Mathes (2004) hypothesised that tag distribution within a set of tags would follow a Zipfian power law distribution and various studies have gone on to confirm this (Guy & Tonkin, 2006; Angus, Thelwall, & Stuart, 2008; Heymann & Garcia-Molina, 2009; Bischoff, Firan, Nejdl, & Paiu, 2008). Ke & Chen (2012) further demonstrated that not only the tag distribution, but also the tag-category distribution, within a folksonomy, "echoed" a power

law distribution for their sample set composed of tags applied to articles on the CiteULike website. The use of the term "echoed" in their study is important, as the y-axis in the tag category usage graph is not calculated on a logarithmic scale, and so it cannot be said that the curve "demonstrates" a power-law curve.

Various researchers discuss the merits of the tags in the long tail, for example Bates & Rowley (2009), as mentioned above, who note that differing "worldviews" can be accommodated well by allowing less frequently-applied tags to exist alongside the more popular ones. According to Shirky (2005), the long tail in a folksonomy can be important to allow individual users to find the resources they need. Shirky does however go on to suggest that in a large scale folksonomy, the full long tail may not be crucial "you try to find ways that the individual sense-making can roll up to something which is of value in aggregate". According to Halpin, Robu, & Shepherd (2007), collaborative tagging within a folksonomy structure tends to move towards a stable set of tags, in that "the tagging eventually settles to a group of tags that describe the resource well and where new users mostly reinforce already present tags in the same frequency as in the stable distribution". This leads them to conclude that in carrying out an analysis of users' tags minus the long tail tags, it should be possible to understand the overall categorisation scheme of the system, that one can "safely ignore the "long-tail" of idiosyncratic and low frequency tags that are used by users to tweak their own results for personal benefit, or alternatively, treat the "long-tail" as an object of examination for other reasons". In a study the following year, Suchanek, Vojnovic, & Gunawardena (2008) discuss the importance of assessing the "meaningfulness" of tags in order to make a good guess at the usefulness of an individual tag for semantic application. "Meaningful" tags according to Suchanek et al. (2008) are tags that identify an item or a characteristic of an item, as opposed to tags that operate organisationally for a user, or are simply "unintelligible" to other users. They found that in general, the more popular a tag was, the more likely it was to have meaning "aggregating the top tags of a document biases to filtering out the meaningful tags", suggesting that in general, the short head of the folksonomy is more useful for information retrieval than the long tail.

2.5 Building category models for tagging systems

A number of different categories to describe the content of users' tags, and the motivations of users when they tag, have been proposed in the literature. Golder & Huberman (2006) analyse two sets of data from website Delicious and analysed user activity, for example the

number of tags used by each user, and also the content of tags. They categorised user tags, with seven main categories of tag emerging as likely to appear on a resource: "identifying what (or who) it is about", "identifying what it is", "identifying who owns it", "refining categories", "identifying qualities or characteristics", "self reference" (tags such as my stuff) and "task organizing". They also discuss the stable pattern that tends to emerge as multiple tags are applied to a resource over time "usually after the first 100 or so bookmarks, each tag's frequency is a nearly fixed proportion of the total frequency of all tags used". Kipp (2007) analysed tags on three internet URL bookmarking sites to examine "the nature and use of non subject tags in tagging systems". Non subject tags, which were found to make up 16% of all tags in Kipp's study, were further broken down into "affective tags" (those indicating emotional response) and "time and task related tags", which, it is proposed, indicates that users have both "an emotional connection to" and "a desire to attach personal information management information to" documents. In a later study, Heckner, Mühlbacher, & Wolff (2008) continued on in the vein of Kipp (2007) by attempting to categorise users' tags in the web-based bibliographic annotation system Connotea. They discussed previous studies on tags and concluded that "in order to provide a reasonable basis for comparison (between user wording and conventional keywording) a category model for existing tags is needed". In their study, content-related keywords are analysed, as are meta-keywords, or keywords that "identify qualities or characteristics beyond mere content description". A further element to the study assessed user tags compared with full text, and found that almost half of all user tags were not found in the document text, thus indicating that users' tags "considerably add to the lexical space of the tagged resource". Following on from the initial 2008 study, Heckner, Neubauer, and Wolff (2008) went on to analyse a larger set of data from four online social tagging websites (del.icio.us.com, flickr.com, connotea.org and youtube.com) and concluded that different resource types tend to be tagged in noticeably different ways. For example, photos tend to be tagged for content, location and device name, whereas scientific articles and web links tend to be tagged with time and task related tags more than other types of content. An eight category model for user tags, comprising "Topic", "Time", "Location", "Type", "Author/Owner", "Opinions/Qualities", "Usage context" and "Self reference", resulted from a study by Bischoff, Firan, Nejdl, & Paiu (2008). The tags studied were applied to websites, images and music, as well as anchor texts from a web crawl. Again, it was noted that categories vary greatly depending on the type of resource being tagged "the distributions of tag types strongly depend on the resources they annotate". The long tail was ignored in all cases "as the long tail consists mostly of idiosyncratic tags with very low usage frequencies,

the influence of this adjustment should be negligible". In their study, Cantador, Konstas, and Joemon (Cantador, Konstas, & Joemon, 2011), similarly to Overell et al, presented a means of automatically filtering raw tags. The focus in this study, however, was on categorisation of the tags themselves rather than on mapping to an ontology. Mapping techniques were used to associate tags on Flickr images with external resources such as Wordnet and Wikipedia, and it was found that tags generally described either "the content of an item", "contextual information about the annotated item", "subjective opinions and qualities" or "self-references and personal tasks". One of the aims of the research was to assess whether this type of categorisation could be useful in moving towards "folksonomy-based recommendation strategies". It was noted that the categorisation of tags can be a difficult process, in part due to misspellings, synonyms, acronyms, morphological derivations, personal assessments and even tags that "are unintelligible to another person". Categorisation was again the focus of Ke & Chen's (2012) study on social tagging of scholarly articles on website CiteULike, which divided the tags into 26 proposed tag categories. They noted that from previous categorisations, it appeared that "the most popularly used category for social tags varies according to the type of tagging objects". Tourné & Godoy's (2012) study attempted automatic analysis of tags applied to web resources. They noted that running tags through a spell check process, and discarding non-matching tags, in order to reduce noise caused by misspellings, caused a loss of information, because the many discarded tags (12%) on further analysis, mostly proved to consist of abbreviations or non-English words. They concluded that both cases should have been considered "to define an enhanced misspelling correction method".

2.6 Understanding the motivation behind user tagging

Understanding the motivations of users is generally at least an indirect focus in the research that aims to build up a category model for tags, with some studies focussing directly on this theme. Within the literature, there is a general consensus that users tend to tag for two main purposes, organisation and communication/description (Ames & Naaman, 2007; Bartley, 2009; Körner, Grahsl, Kern, & Strohmaier, 2010).

In their 2006 study, Marlow, Naaman, Boyd, & Davis (2006) discussed tagging systems and how users' motivations in tagging affect them "the personal and social incentives that prompt individuals to participate affect the system itself in various ways". They discussed the variance between users, some of whom tag for themselves, others of

whom have the group in mind, still others a combination of the two "many users begin with the conception that they are tagging for themselves; some begin to appreciate the sociable aspects over time, while others have no interest in that component". Motivations ranging from "future retrieval" to "opinion expression" were discussed, and they suggested that the types of tags found in a system can be viewed as the result of the users' motivations being expressed through their tags. The separation between individual use and collective use is again highlighted by Guy & Tonkin (2006), who assert that extensive personal tag use may reflect the "real problem with folksonomies.. that they are trying to serve two masters at once; the personal collection, and the collective collection". Morrison (2007) recommends that tagging system designers should take into account users' motivations for tagging "a folksonomy is more likely to be successful when the goals of the website or information system intersect with the goals and motivations of users". He also defines some general user motivations for tagging, such as "future retrieval", with more specific motivations, such as "to play a game or earn points" depending on the tagging system and the type of resources being tagged. In an interview-based study, Ames & Naaman (2007) discovered that users' motivations for adding tags to online resources were generally for organisation and communication, and these two main motivation areas were further divided by whether the organisation or communication was for social or for personal purposes. Bartley's (2009) study attempted "to understand book tagging by investigating LibraryThing (LT) members' purposes for tagging". Questionnaires were distributed to members about the reasons they apply tags, with results showing that 74% of users indicated "collection management" was their primary reason for tagging, followed by "recording factual information" and "helping others find the book". Lu, Park, & Hu, (2010) discussed the importance of analysing user motivation in tagging systems "user-created tags provide a window into users' interests, behaviours and attitudes that might help information institutions better understand and server users". Similarly to Ames & Naaman (2007), Körner, Grahsl, Kern, & Strohmaier (2010) asserted that users have two main motivations when they tag – categorisation, to allow them to "construct and maintain a navigational aid to the resources for later browsing" (organisation) and description, to allow them to "accurately and precisely describe resources" (communication). Social/personal tagging was also discussed in this study, which proposed that motivations fall into the two main areas, organisation and communication, and that these are further subdivided into that which is undertaken for the "self" or for the "group".

2.7 Research specifically on book tagging

Although the tagging on other resources such as web documents and images is more common in the research, tagging on books, and in particular on LibraryThing tags, has had some analysis and discussion in the literature. The original research and datasets generated for this study provide further source material as a contribution to the evolving discussion. This will be considered in greater depth in Chapters 5 and 6, "Discussion" and "Conclusion".

In his discussion on the benefits of "leveraging communities" in order to improve folksonomies, Smith (2008) mentions the LibraryThing website's feature allowing users to make any two tags equivalent, thus forming clusters of tags with the most popular tag being the "preferred term". As he mentions, the only tags for which this generally can be considered not to involve information loss are pairs of tags "where the sociosemantic delta is zero" such as "World War 2" and "WWII". He notes that the community tends to be conservative in these pairings, keeping, for example "humor" and "humour" separate due to the value of having what is considered humorous in the US differentiated from what is considered humourous in the UK. Lawson (2009) discussed tagging on Amazon.com and LibraryThing and assessed the quality of social tagging and its comparison with Library of Congress subject headings for similar content. Lawson found that "subjective" book tags, or those that do not deal with the content of the book, generally fall into twelve main categories: "Reading Status", "Date", "Initials of tagger", "Type", "Gift suggestion", "Format", "Referral", "Location", "Bibliographic", "Opinion", "Author" and "Publisher". Bartley (2009) as mentioned above, set about investigating the motivations of LibraryThing members for adding tags to books. Thomas, Caudle, & Schmitz (2010) also took LibraryThing as an example, taking ten books and analysing all the tags (a total of 7653) on those books to ascertain the percentage of "messy" tags that tend to be included in book folksonomies. They defined messy tags as tags that "affect general search and retrieval because of the variation among tags", including tags that include nonalphabetical characters or dates, variations of other tags, foreign language tags and misspellings. A decision was made to discard all personal tags (for example variations on the verbs read and own and on the nouns box and shelf) for the purposes of the research. Another study on LibraryThing tags, carried out by Iyer & Bungo (2011), analysed forty books and qualitatively analysed them for matches between user tags and subject headings (mostly Library of Congress subject headings) from their associated MARC records in the OCLC database. The categorisation of the remaining

tags that did not match subject headings was deemed necessary as "individual tags do not lend themselves to semantic analysis because they vary so widely, they do not have context when they stand alone and there are simply so many of them. When grouped by conceptual similarity, rather than alphabetically or by frequency, as in tag clouds, the context becomes richer and more meaningful". Again, confirming the experiences of previous studies, Iyer & Bungo came across some tags that were simply "undecipherable".

2.8 Summary

Tagging and folksonomies in general began to be studied in earnest from about 2004 on. Although there have been a number of studies categorising user tags on various systems, the literature is somewhat sparse on analyses of book tagging, with regard to the categories into which tags fall, and the motivations that this might reveal.

Having reviewed the available literature, it became clear that many studies did not take into account the frequency of application of a tag to a particular resource when categorising tags. This is crucial information, as it allows tags to be weighted according to their popularity with users, and thus provides for a more accurate category model. Therefore, one important part of the research question became the generation of a booksonomy category model with this frequency data included.

Confirming the power law distribution of tags, while not a main feature of the study, was also decided on as a useful undertaking given such a large set of tag and frequency information. As will be seen in the study results detailed in Chapter 4, the insight generated by combining both tag and frequency information enhances the literature on distribution of tags, and can help to clarify motivation for tagging, particularly in comparison to other studies which have not disaggregated these two variables.

Users' motivations for tagging are generally considered by the literature to fall into two main areas, organisational and subject-based, with the reasons for adding tags within these two areas again sub-divided into more personal motivations and more social motivations. An assessment of whether the motivations of book taggers falls in with this general model, based on the arrived-at category model, was decided on as a further useful component of the research question.

3 Methodology

3.1 Introduction

An overview of the research strategy and the methodology followed is provided in this chapter. The reason why the particular research strategy was chosen is included. The processes of data gathering, data normalisation and analysis methods are also detailed.

3.2 Choosing a research method

A quantitative approach was the chosen research method, due to the requirement to generate a meaningful category model across a large number of tags. Having large amounts of tag and category data would allow for statistical analysis and enable some extrapolation to be made in discussions about book tags as a whole.

3.3 Choosing the data

Having studied a number of tagged books, it was decided that an analysis of fifty popular books would provide a statistically valid and academically useful sample of tags for analysis. In order to ensure a rich selection of tags, it was important to choose books that were popular on the site and so would have a good rate of tagging associated with them. The LibraryThing administrators provide a list of the 250 most reviewed books (LibraryThing, 2012), which could also be assumed to be quite highly tagged as they were of interest to many users. The list of top reviewed books was pulled from the LibraryThing website in early September 2012, and in order to make the selection random within this group, every fifth book of this list was selected, resulting in a group of fifty books to be analysed (see Appendix A: List of books for analysis). Where there were multiple editions of a book available on the site, the edition most commonly chosen by members was selected, again to ensure high levels of tagging.

3.4 Should the long tail be analysed?

A decision had to be made as to how far down the frequency list the categorisation should go for each book. In order to cover a reasonably representative sample of books, it was determined that categorising all tags for each book would not be possible due to resource constraints. Based on a thorough analysis of the tag counts and frequencies for a smaller sample set of three books, it was concluded that taking tag frequency into account as well as tag count and by limiting the analysis to tags with frequency 3 and above, the study would still cover at least 80% of the data. It was therefore decided that a cut off of frequency 2 or

less would be used to decide on the "long tail" that would not be categorised. This decision to cut off some of the long tail was based on the literature (Halpin, Robu, & Shepherd, 2007; Suchanek, Vojnovic, & Gunawardena, 2008). In practice, the average percentage of the data covered for the fifty books, taking tag frequency into account, was 83%.

3.5 Retrieving the tags

Tags were retrieved during a one week period between the 8th and 15th of September 2012. The tags for each book were retrieved in HTML format using the Firefox browser (Figure 3-1), and then pasted into Microsoft Word in "Unformatted Unicode Text" format (Figure 3-2).

Figure 3-1 A partial view of tags for book "The Hunger Games"

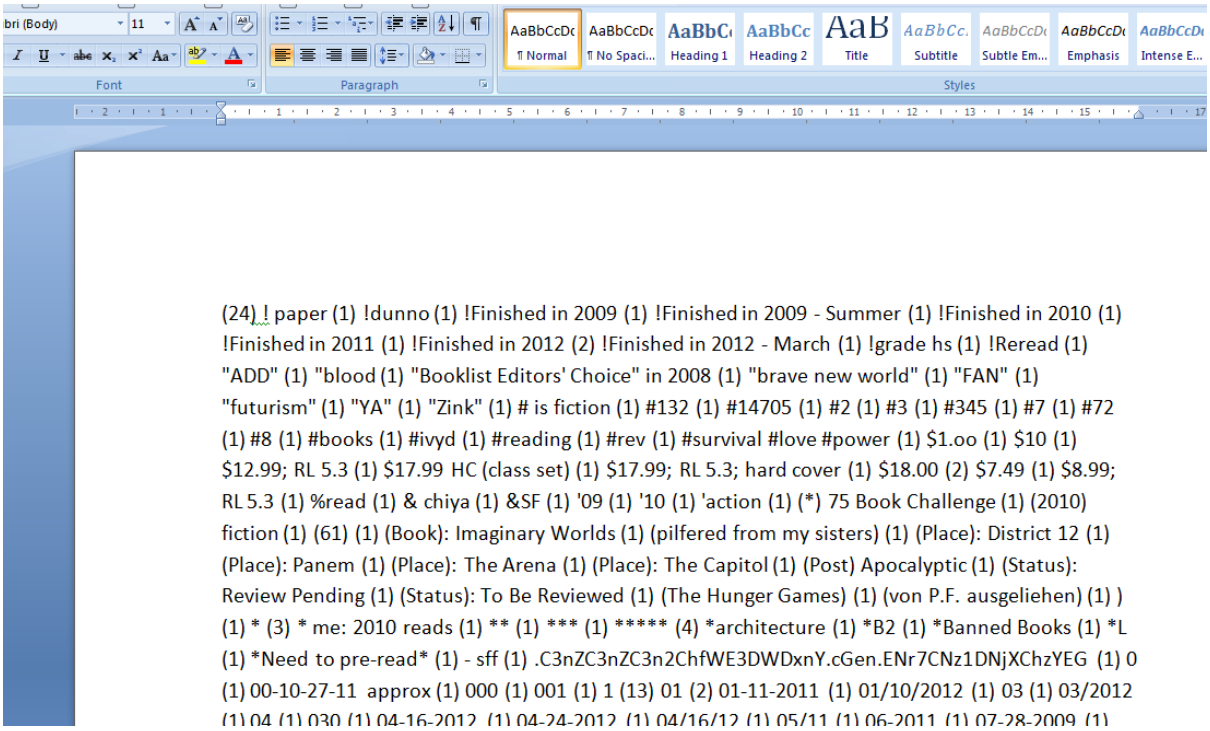

Figure 3-2 Tags pasted into Microsoft Word in "Unformatted Unicode Text" format

In order to form a list of tags in a clean enough format for analysis, Word's "Replace All" function was used to replace all closing parentheses ')' with the special character string '^p' which resulted in one tag per line in the Word file. This data was then copied and pasted into a Microsoft Excel spreadsheet.

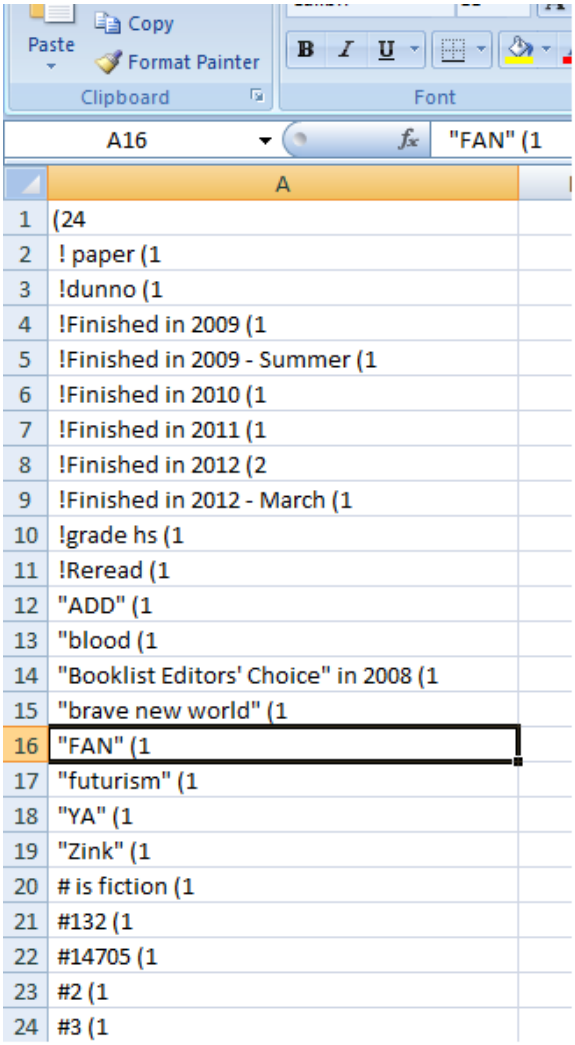

Figure 3-3 Initial tag list in Microsoft Excel

3.6 Cleaning the data

The initial step necessitated separating the tag itself from its frequency information. A macro was designed and developed to do this; applying "delimited" formatting to separate each line of text at an opening bracket, giving two columns of data, a "Tag" column and a "Frequency" column.

A further cleanup step was then required to ensure that any tags that themselves contained brackets, would be reinstated correctly. This was done by applying a macro that used "conditional formatting" to highlight non-number text (see 3.8.2) in the "frequency" column and reinstating the tag and its associated frequency manually. Although many of these bracket-containing tags when cleaned up, ended up being long tail tags of frequency 1 and 2 (see 3.4), it was still deemed necessary to carry out this long-handed clean up in order to make sure not to inadvertently exclude any important non-long-tail tags from the analysis.

In some cases, users had added tags that consisted of copies of lists of multiple tags and their frequencies, as a single tag. To combat this, a further macro had to be applied to the "Tags" column to highlight duplicates, and the invalid duplicates were then removed. See Figure 3-4 for an example of this.

Figure 3-4 List of multiple tags pasted as one tag

Another issue that arose was tags that contained a number within them in parentheses, for example *Read(2010)*. These tags, if not cleaned correctly, could cause incorrect frequency data, as the tag would be incorrectly stored as the text before the brackets (rather than the full text) and the number (in this case a year) would be interpreted as the frequency of that tag. A further conditional formatting macro (see 3.8.2) was thus applied to the "Frequency" column to find all frequencies greater than 1000. In some cases, these were valid frequencies, but it was generally very clear when they were actually referring to years, and the tag could thus be amended to give it the correct frequency.

Once all tags had been normalised, the list of tags was sorted by descending frequency. Two extra columns "Category" and "Notes" were added, and the Category column was set up with Data Validation to allow only items from a separate "Categories" tab (see Appendix C: List of possible categories) to be entered from dropdowns. The categorisation was then carried out for each tag, with a separate list for each book (see for example Figure 3-5). The "Notes" column was used to make extra notes on a categorisation, for example to mark where tags involved abbreviations, to mark what language an "other language" tag was in, and so on.

	A	B	C	D
1	Tag	Frequency	Category	Notes
2	fiction	4,702	genre/style	
3	mystery	2,123	genre/style	
4	thriller	1,769	genre/style	
5	religion	1,146	subject	
6	suspense	728	genre/style	
7	read	702	personal task-based	
8	novel	608	genre/style	
9	conspiracy	503	subject	
10	Leonardo da Vinci	434	subject	
11	art	411	subject	
12	adventure	340	genre/style	
13	Dan Brown	308	author information	
14	Christianity	291	subject	
15	history	289	subject	
16	Holy Grail	262	subject	
17	France	230	location	
18	historical fiction	223	genre/style	
19	Paris	213	location	
20	own	196	personal task-based	
21	Mary Magdalene	183	subject	
22	Robert Langdon	179	character/setting information	
23	Knights Templar	177	subject	
24	crime	174	subject	
25	grail	162	subject	
26	secret societies	154	subject	
27	Opus Dei	144	subject	
28	movie	137	movie information	

Figure 3-5 Tags in Excel with categorisation

3.7 Categorising the data

3.7.1 Initial categorisation trial

It was of course very important that categorisation be consistent across all fifty books in order for the final results to have meaning. Therefore, in order to set up a final model for the categorisation, the tags on five books were initially analysed to address early as many as possible of the categorisation questions that might arise in later stages of the study. An example of this initial categorisation model can be seen in Appendix D: Initial categorisation example, which displays the initial and revised categorisation results for one of the initial five books categorised.

3.7.1.1 First revisions of the category model and categorisation process

After the initial subset of five books had been categorised, the resulting categories were assessed for their usefulness and the practicality of using them to categorise. At this point, a decision was made to streamline and remove some very specific categories in order to allow for a broader and clearer summarisation of the final data. For example, initially the category "action" was used to categorise tags such as *to read* and *loaned to Susan* and similar, and the "physical item" category was used for tags like *in box in attic, owned* as well as the more obviously physical tags like *hardback, blue cover*. The "action" category was removed, it was decided that "physical item" should now only be used for specifically physical

characteristics of the actual book "object" itself, and a new broader "personal task-based" category was created to encompass the above different types of personal tag. Location categories were also streamlined, with two location categories ("location of book" and "location author information") becoming just "location". Where a location tag was determined to be referring to the author, the category "author information" was used. As a final example, the two initial subject categories ("subject – specific" and "subject-general") were merged into one general "subject" category.

One situation that arose several times was a tag having two applicable categories – for example *read in 2007* (task-based: *read in*; date: *2007*) or young adult fiction (target reader: *young adult*; genre: *fiction*). In cases like this a decision was made on an individual tag basis as to which was the more important category for the tag (in the first given example, a category of "personal task-based" would have been assigned, and in the second, a category of "target reader" would have been assigned). A record was kept of these types of decisions and when similar situations arose, this record was referred to and a parallel decision was made.

It was also decided that misspellings including word formatting errors (for example words running together with no spaces) would be marked in a separate column so that these could be analysed separately.

3.7.1.2 Later revisions

Not all tag scenarios had arisen or become apparent as representing a pattern/category during the initial five book trial, so some revisions had also to be made throughout the categorisation process. For example, on working through several books targeted at young adults, it became clear that a "target reader" category was needed, as was a "reading system" category.

Further categories added after the initial trial were "translator/narrator/illustrator" and "publisher information" as tags came up that required these categorisations and could not fit into any of the initial categories. Indeed, "translator/narrator/illustrator" itself evolved from the initially added "translator information" as further books introduced tags based on their narrators and illustrators, and having three categories was deemed overly specific.

3.7.2 General categorisation method

Where a tag was difficult to place in a category, a ? category was assigned and the tag was revisited at a later time with any other tags in the ? category for that book. The Wikipedia website (Wikipedia, 2012) was very useful in retrieving information about characters and

settings of specific books. At other times, a Google search (Google Inc., 2012) with the book title, author name and the tag was used in order to understand the meaning of a tag and thus its category. The AcronymFinder website (Acronym Finder, 2012) was also used to decipher tags that were sequences of letters.

3.7.3 Category explanations and tag examples

Specific tag examples are shown in italics. Numbers in parentheses after tags show the frequency of application of the tag.

3.7.3.1 Category: ?

Description: Non-code, natural language tags that could not be categorised
Examples: *Cathars (3), i think L'll Go Flying (8), kolzow (4), Torney (42)*

3.7.3.2 Category: Author information

Description: Information about the author, for example gender, nationality, death
Examples: *American author (64), Diane Setterfield (34), female author (102), posthumously published (3)*

3.7.3.3 Category: Awards/popularity

Description: Information about any awards won or the popularity overall of a book.
Examples: *1001 Books to Read Before You Die (200), award winner (141), Newbery (745), popular fiction (44)*

3.7.3.4 Category: Blank

Description: A blank tag occurred in the data for almost every book. It is not clear how these came to be in the data set, as attempts made to deliberately add a blank tag to a book failed. Blank tags were included in the quantitative but not the qualitative elements of the analysis.

3.7.3.5 Category: Character/setting information

Description: References to characters or fictional settings of books.
Examples: *female protagonist (96), Gollum (51)*
Additional note: The category "character/setting information" was only used for locations where they were fictional (for example *Camp Half-Blood*: The Lightning Thief) but not for non-fictional setting locations (for example *Pacific Ocean*: Life of Pi).

3.7.3.6 Category: Code

Description: Tags not in natural language that could not be categorised elsewhere. (Note: where possible, abbreviations were assigned to the appropriate category (for example F was categorised as genre/style as an abbreviation of fiction).

Examples: *@Woking_S1_6 (26), MG (24), bab (15), SI624fall10 (8)*

Additional note: Not all "codes" were put in the "code" category – for example *F* is a standard usage for Fiction so this tag was given category "genre/style".

3.7.3.7 Category: Date

Description: Any date referenced, including time periods. No distinction is made between the date a book was published, for example, and the date in which it is set.

Examples: *1970s (50), Middle Ages (380), 19th century (2280), 2007 (823)*

Additional note: Generally, where a date was mentioned, the "date" category was applied (for example *19th century literature*). The "date" category was also used for named periods (for example *Regency*, *Victorian England*).

3.7.3.8 Category: Genre/Style

Description: Referring to the literary form, technique or style of the book.

Examples: *non-fiction (5157), steampunk (460), allegory (112)*

3.7.3.9 Category: Language of book

Description: Any reference to a language was assumed to refer to the language of the book.

Examples: *German (110), Language: English (20)*

3.7.3.10 Category: Location

Description: Any location. Note that completely fictional locations were given the category character/setting information.

Examples: *Paris (366), China (747)*

Additional note: Location was only used for geographic locations. For other locations, subject was used (for example *Louvre*: The Da Vinci Code).

3.7.3.11 Category: Movie Information

Description: References to a film or film series based on the book or series of which it is a part.

Examples: *film adaptation (61), Tom Hanks (6)*

3.7.3.12 Category: Opinion

Description: References to an emotion or opinion about the book.

Examples: *made me cry (17), overrated (70), loved (43)*

3.7.3.13 Category: Other language

Description: Any tag not in English. The notes column was used to specify the language of the tag (Google Translate's "Detect language" functionality was useful in this).

Examples: *skönlitteratur (226), fantastique (53), literatura estrangeira (7)*

3.7.3.14 Category: Personal task-based

Description: Tags that are assumed to refer to actions that have been or will be taken by the person tagging.

Examples: *already read (563), unfinished (111), mom (10)*

Additional note: Where a name was mentioned that was not either the author's name, a character's name, or a name that could be found to be associated with the book, the tag was coded personal task-based.

3.7.3.15 Category: Physical item

Description: Refers to the physical characteristics of the book itself.

Examples: *Kindle (1030), leather bound (44), Large Print (21)*

3.7.3.16 Category: Publisher information

Description: Information about the publisher of the book.

Examples: *Easton Press (110), Everyman's Library (76).*

3.7.3.17 Category: Reading system

Description: References to systems that apply points or grades to books in order to help readers to assess the difficulty of the book or to monitor their reading (e.g. "Accelerated Reader").

Examples: *AR 4.6 (15), Sonlight 5 (9), Level R (26)*

3.7.3.18 Category: Reference

Description: A small category, but one that was needed where a tag referred to another work of literature, for example.

Examples: *Romeo and Juliet (14), William of Baskerville (9)*

3.7.3.19 Category: Series information

Description: Where a book is one of a series, tags referring to its place in the series, to the series name, etc.

Examples: *Twilight saga (214), prequel (56), Trylle trilogy (14)*

3.7.3.20 Category: Subject

Description: Tags that refer to what the book is about, either specifically or more generally.

Examples: *friendship (2546), atheism (1238), footbinding (275), HeLa (64)*

3.7.3.21 Category: Target reader

Description: Information about the type of reader the book is primarily aimed at.

Examples: *young adult fiction (807), jfic (81), tween (32)*

Additional note: Where there was a question between subject/target reader, target reader was generally chosen rather than subject (for example *teenagers*: New Moon). This decision was made based on the general proliferation of "target reader" category tags for these types of books. Also, where school grades were mentioned, these were given the "target reader" category (for example 6^{th} *grade,* 7^{th} *grade*: The Lightning Thief).

3.7.3.22 Category: Title information

Description: Tags referencing the title of the book.

Examples: *The Hobbit (52), eyre (15), Tuesdays (5)*

3.7.3.23 Category: Translator/narrator/illustrator

Description: Information about the translator, the narrator or the illustrator of a particular book.

Examples: *Pevear and Volokhonsky (5), Stephen Fry (3)*

3.7.3.24 Category: Website

Description: References to websites.

Examples: *bookcrossing (70), audible.com (4)*

3.8 Summarising and analysing the information

The Microsoft Windows application Excel 2007 was used to generate the statistics set out in detail in Chapter 4 "Results".

3.8.1 Pivot tables

Pivot tables were used to sum up the categorisation information for all books. Pivot tables are a means of generating data summaries based on dynamically-chosen features of a given data set. They allow for the "rolling-up" and display of multi-faceted data based on dynamically-chosen facets.

Specifically, the pivot tables were generated from the full set of data, including book reference number, tag, frequency, category and notes information for each book (see Figure 3-6). Various pivot tables were built from this data, depending on the focus of the summarisation needed. To generate the full tag count and frequency data for all books, for example, the pivot table was set up as shown in Figure 3-7. The resulting pivot table, with one category expanded to show how the pivot table hierarchy functions, is shown in Figure 3-8.

	A	B	C	D	E	F
1	Book	Tag	Frequency	Category	Notes	
236	1	power of words	6	subject		
237	1	realistic fiction	6	genre/style		
238	1	reviewed	6	personal task-based		
239	1	school	6	personal task-based		
240	1	Your library	6	personal task-based		
241	1	2005	5	date		
242	1	basement	5	personal task-based		
243	1	BBYA	5	awards/popularity	ALA Best books for young adults	
244	1	beautiful	5	opinion		
245	1	chapter book	5	genre/style		
246	1	crossover	5	genre/style		
247	1	Death as a character	5	character/setting information		
248	1	female protagonist	5	character/setting information		
249	1	foster families	5	subject		
250	1	HC	5	code		
251	1	jødeforfølgelse	5	other language	Norwegian	
252	1	lezen	5	other language	Dutch	
253	1	male author	5	author information		
254	1	Markus	5	author information		
255	1	Molching	5	location		
256	1	narrator	5	genre/style		
257	1	National Jewish Book Award	5	awards/popularity		

Figure 3-6 Example view of data

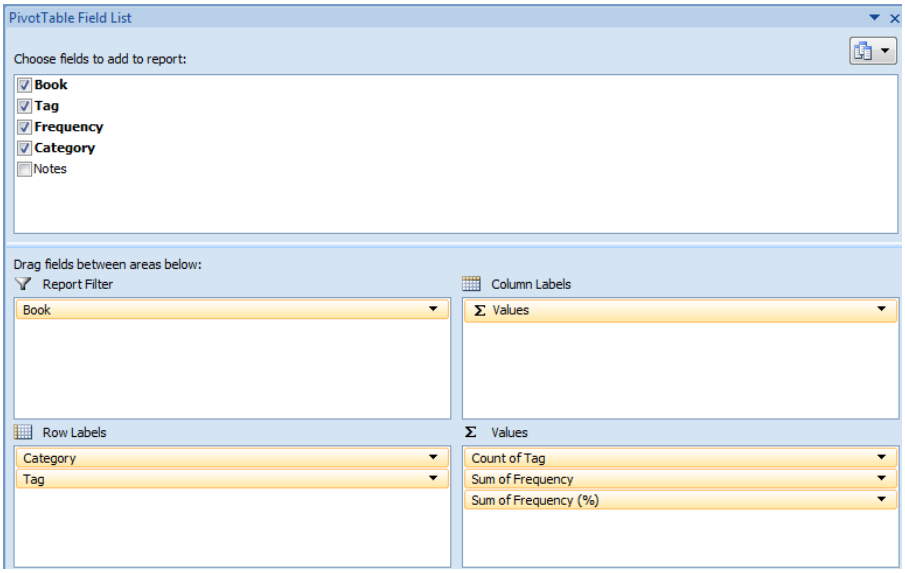

Figure 3-7 Pivot table set-up example

	A	B	C	D	E	F
1	Book	(All)				
2						
3		Values				
4	Row Labels	Tag Count	Tag Count (%)	Sum of Frequency	Sum of Frequency (%)	
5	genre/style	1487	11.16%	158591	36.2%	
6	subject	4086	30.67%	120981	27.6%	
7	personal task-based	2009	15.08%	32782	7.5%	
8	target reader	675	5.07%	23653	5.4%	
9	location	353	2.65%	17958	4.1%	
10	author information	411	3.09%	16423	3.7%	
11	date	768	5.77%	15228	3.5%	
12	physical item	690	5.18%	9317	2.1%	
13	awards/popularity	320	2.40%	6940	1.6%	
14	character/setting information	301	2.26%	6727	1.5%	
15	opinion	520	3.90%	5599	1.3%	
16	series information	112	0.84%	5266	1.2%	
17	title information	78	0.59%	4370	1.0%	
18	code	547	4.11%	3189	0.7%	
19	other language	357	2.68%	2942	0.7%	
20	language of book	158	1.19%	2344	0.5%	
21	movie information	97	0.73%	1817	0.4%	
22	blank	8	0.06%	1570	0.4%	
23	publisher information	80	0.60%	973	0.2%	
24	reading system	105	0.79%	843	0.2%	
25	website	103	0.77%	514	0.1%	
26	?	33	0.25%	191	0.0%	
27	reference	12	0.09%	65	0.0%	
28	translator/narrator/illustrator	11	0.08%	57	0.0%	
29	Alan Lee	1	0.01%	13	0.0%	
30	Jim Dale	1	0.01%	8	0.0%	
31	Larissa Volokhonsky	1	0.01%	6	0.0%	
32	Pevear-Volokhonsky Translation	1	0.01%	5	0.0%	
33	Pevear & Volokhonsky	1	0.01%	5	0.0%	
34	Richard Pevear	1	0.01%	4	0.0%	
35	Michael Hague	1	0.01%	4	0.0%	
36	william weaver	1	0.01%	3	0.0%	
37	Stephen Fry	1	0.01%	3	0.0%	
38	Hague	1	0.01%	3	0.0%	
39	Pevear / Volokhonsky	1	0.01%	3	0.0%	
40	Grand Total	13321	100.00%	438340	100.0%	
41						

Figure 3-8 Pivot table example

Setting the filter on the "Book" data field allowed for the filtering of the data based on specific books, which was the method used to generate statistics for specific book types (see section 4.4.5).

3.8.2 Excel formulae

In all cases, B2 is the Excel cell containing the tag.

Highlighting non-number text in the "frequency" column was necessary in order to find any tags that included brackets. Highlighted tags were then reinstated manually.

=ISNUMBER(B2)=FALSE

Some tags contained years within them in brackets, causing the macro to format them as though the year was the frequency. An additional conditional formatting formula was applied in order to highlight these tags for manual cleanup:

=IF(B2 > 1000, TRUE, FALSE)

Calculating tag length:

=LEN(TRIM(B2)) where B2 is an Excel cell containing the tag.

Calculating the number of words per tag:

=LEN(TRIM(B2))-LEN(TRIM(SUBSTITUTE(B2," ","")))+1

3.9 Limitations

The quantitative nature of the study, while informative and soundly constructed, leaves some questions to be answered. Categorising tags without having the exact meaning of the tag explained, is difficult, and it was necessary that some assumptions be made. If a qualitative portion of research had been included in the study (for example interviewing some taggers who had applied tags to the books) it might have been more helpful towards the categorisation itself, but also towards understanding the motivations of those users, and thus linking the users' tags, and also similar tags of other users, with the category model. Similar studies within the literature, but that contained a qualitative as well as a quantitative aspect (Ames & Naaman, 2007; Bartley, 2009), are considered, on reflection, to provide more context and a more rounded view of tagging practice. This type of qualitative data would have been difficult to retrieve for this particular set of data, however, and so adding this

approach might have limited the study greatly in terms of what the quantitative portion could have covered.

The fact that only fifty books could be analysed with respect to their applied tags, and that resources were not available to analyse the tags within the "long tail" is also a limitation of the study. Furthermore, as the categorisation was carried out by only one person, there is necessarily quite a high degree of subjectivity involved, with no opportunity for cross-referencing of categories applied.

3.10 Summary

This chapter has outlined the research strategy chosen and the methodology used within the study. A quantitative approach was chosen to allow for statistical analysis and some extrapolation from the findings about the sample set to booksonomies as a whole. The data was gathered over a period of a week, and fifty books out of the 250 from the top reviewed list on the LibraryThing website were chosen for analysis, due to the fact that higher reviewed books tend to have a greater number of tags applied. Data was passed through several processes in order to clean and normalise it, with macros, formulas and manual cleanup steps used where appropriate. The categorisation process was a two step one, with an initial sample set of five books categorised in order to build a robust category model of twenty four categories. This model was then used in order to categorise the full set of fifty books. Pivot tables were used within the Microsoft Excel software package in order to analyse and display the data as a multiple-level hierarchy with amalgamated totals for various statistics. One limitation of the methodology is the fact that no contextual information was available for tags, which sometimes made it difficult to assign tags to a particular category with assurance. As well as this, categorising the tags was necessarily subjective, as only one person was involved in the categorisation process.

4 Results

4.1 Introduction

There are two main components of the results data - the tags themselves and their associated statistics, and the data resulting from the categorisation of those tags. The two groups of data are presented in separate sections within this chapter.

4.2 Statistics

The total number of tags on all fifty books was 95,134. Of these, 13,358 tags were analysed and categorised. These represented the tags that had a frequency of application of 3 or greater. Taking the frequency of application of the tags into account gives a figure of 438,340 for all analysed tags, and 528,826 for the full set of tags for the fifty books (see Appendix B: Statistics for the set of analysed books). This means that the 13,358 tags were applied an average of 39 times each (of course the actual frequency of application varied widely across tags).

4.3 Tag data

4.3.1 High frequency tags

The top ten tags across the fifty books are shown in Table 4-1. Tag count reflects the number of books on which the tag was applied.

Tag	Category	Tag Count	Sum of Frequency	Sum of Frequency (%)
fiction	Genre/style	49	54190	12.4%
fantasy	Genre/style	29	25657	5.9%
young adult	Target reader	32	9460	2.2%
read	Personal task-based	50	8439	1.9%
novel	Genre/style	48	7007	1.6%
mystery	Genre/style	30	6468	1.5%
classic	Genre/style	19	5716	1.3%
humor	Genre/style	24	5541	1.3%
non-fiction	Genre/style	23	5157	1.2%
religion	Subject	22	5114	1.2%

Table 4-1 Most frequent tags

4.3.2 The long tail of tag data

For the fifty studied books, the average number of total tags per book was 1903, and the average number of tags with a frequency of greater than 2 was 267 – or about 14%. When tag frequency was taken into account, however, the percentage changed hugely, meaning that

analysing all "non-long-tail" tags (where the long tail was decided to have a cut off of 2 or lower) covered 83% of the total tag applications.

Chart 4-1 displays tags plotted against the log of their frequency. The resulting distribution demonstrates a typical Zipfian power-law pattern.

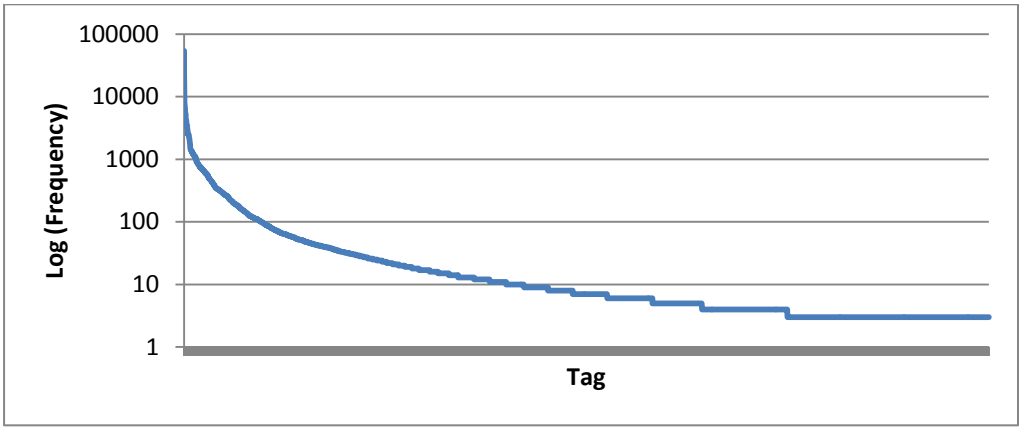

Chart 4-1 Tags plotted against log of tag frequency

4.3.3 Number of words per tag

Data was analysed to assess the number of words per tag, and this information was summarised using a pivot table. The highest number of words in a single tag was 10, with the most common being 1 word, at 63%.

Number of Words in Tag	Tag Count	Tag Count (%)	Tag Count & Freq.	Tag Count & Freq. (%)
1	8412	63.0%	342097	78.0%
2	3706	27.7%	80608	18.4%
3	1064	8.0%	13411	3.1%
4	120	0.9%	1615	0.4%
5	29	0.2%	327	0.1%
7	10	0.1%	211	0.0%
6	12	0.1%	45	0.0%
8	2	0.0%	17	0.0%
10	2	0.0%	6	0.0%
9	1	0.0%	3	0.0%
Grand Total	**13358**	**100.0%**	**438340.1**	**100.0%**

Table 4-2 Number of words per tag

4.3.4 Tag length

Patterns in the length of individual tags were also investigated. Figure 4-1 displays a frequency distribution showing the number of tags of each given length. The shortest tags

were 1 character in length (e.g. *A*, *X*, *9*), while the longest tag was 60 characters in length (*Hogwarts School of Witchcraft and Wizardry (Imaginary place)*).

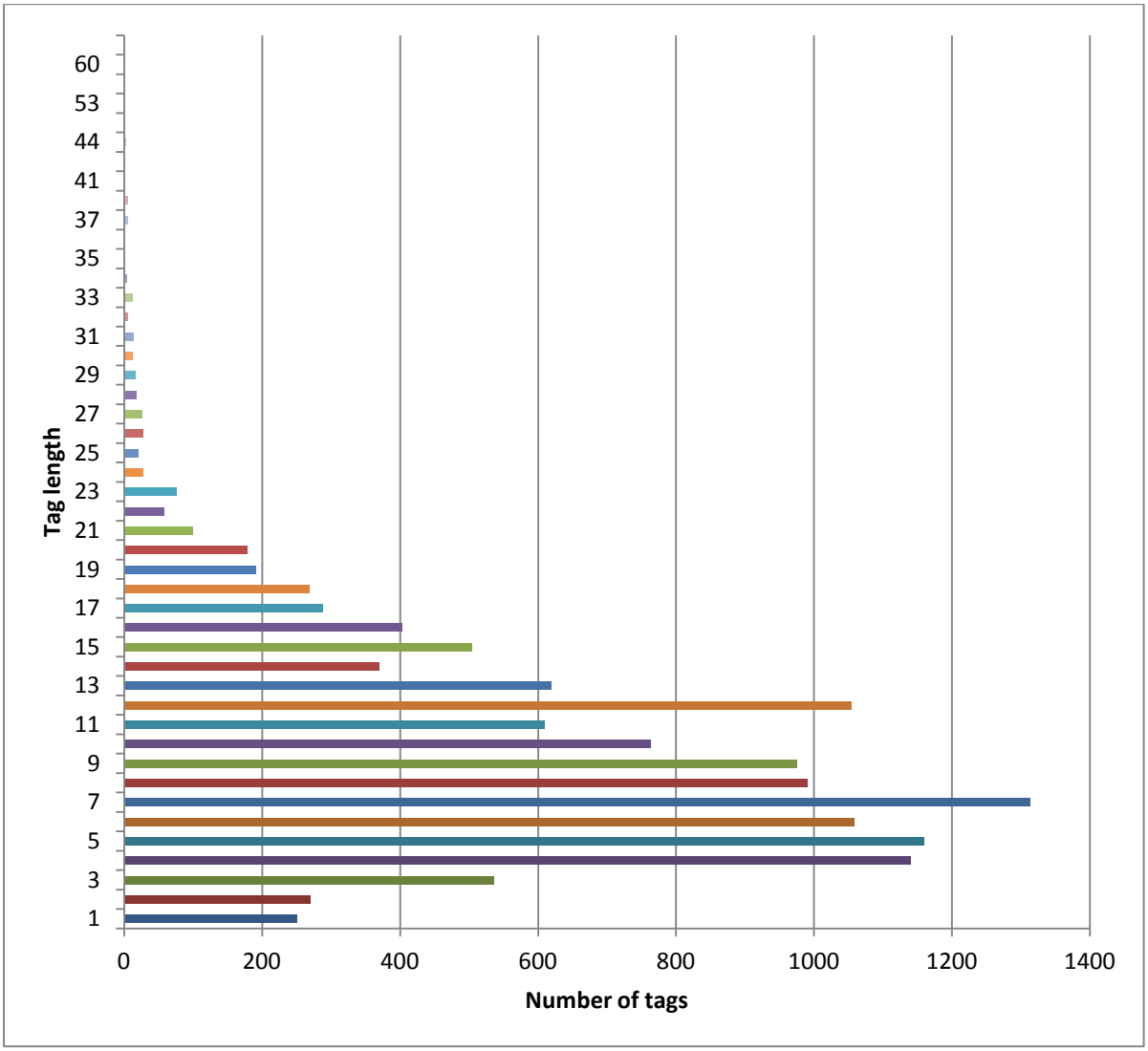

Figure 4-1 Frequency distribution of tag length

Tag Length	Tag Count	Tag Count (%)	Tag Count & Frequency	Tag Count & Frequency (%)
7	1313	9.8%	131907	30.1%
5	1159	8.7%	39864	9.1%
8	989	7.4%	33026	7.5%
6	1058	7.9%	32304	7.4%
4	1140	8.5%	31232	7.1%
9	975	7.3%	26613	6.1%
11	609	4.6%	23875	5.5%
10	763	5.7%	23139	5.3%
12	1054	7.9%	22976	5.2%
13	619	4.6%	12151	2.8%

Table 4-3 Top 10 tag lengths

4.3.5 Miscellaneous interesting tags

There were a number of tags that stood out as being either neologisms or unlikely to appear in a formal ontology. Some examples are: *Multiple plots*, *bibliomystery*, *book within a book*, *magical realism*, *Unreliable narrator*, *Metafiction*, *surprise ending*, *Robinsonade*, *Reluctant readers*, *Read out loud, read to my kids*, *Irreligion*, *Airplane reading*. The appearance of this type of useful, non-formal tag demonstrates the value of the folksonomy compared with a formal ontology, as discussed in the literature.

Several amusing tags were also noted during the course of data retrieval, cleanup and categorisation, including: *used as toilet paper during the Morocco trip (360 pp left)*: Da Vinci Code); *crime (against literature!)*: Da Vinci Code; *Meh*: Running with Scissors.

4.3.6 "Other language" tags

In total, "other language" tags made up 0.7% of the dataset. Table 4-4 shows the breakdown of tags for each of the other represented languages. French was the most prevalent, at 35.2%, followed by German at 22.4%, with Dutch, Swedish and Italian all similarly popular at 9.3%, 9.0% and 8.7% respectively.

Language	Tag Count	Tag Count & Freq.	Tag Count & Freq. (% of Other Language Tags)	Tag Count & Freq. (% of All Tags)
French	62	1036	35.2%	0.25%
German	93	658	22.4%	0.15%
Dutch	43	274	9.3%	0.06%
Swedish	30	266	9.0%	0.06%
Italian	36	257	8.7%	0.06%
Finnish	41	184	6.3%	0.04%
Spanish	28	147	5.0%	0.04%
Norwegian	7	58	2.0%	0.01%
Portuguese	4	15	0.5%	Negligible
Polish	3	11	0.4%	Negligible
Czech	1	7	0.2%	Negligible
Danish	2	6	0.2%	Negligible
Indonesian	1	5	0.2%	Negligible
Hungarian	1	3	0.1%	Negligible

Slovenian	1	3	0.1%	Negligible
Turkish	1	3	0.1%	Negligible
Lithuanian	1	3	0.1%	Negligible
Russian	1	3	0.1%	Negligible
Japanese	1	3	0.1%	Negligible
Total	**357**	**2942**	**100.0%**	**0.7%**

Table 4-4 "Other language" tag statistics

4.3.7 Misspellings

A total of 16 tags were marked as misspelled (total frequency of 99), a percentage of only 0.02% of all tags by frequency.

4.4 Categorisation data

4.4.1 Analysis details

Categorisation data was analysed in two main ways:

1. A simple count of the number of tags in a given category. So, for example, if on one book the tag *fiction* had a frequency of 950, and the tag *to be read* had a frequency of 20, each would only add 1 to the total for their assigned categories ("genre/style" and "personal task-based").
2. A combination of the count and the frequency of each tag (this gives a more accurate view of the "importance" of each tag and thus weights the addition of the tag to the category into which it is categorised. So, for example, if on one book the tag *fiction* had a frequency of 950, and the tag *to be read* had a frequency of 20, the former would add 950 to the total for its assigned category ("genre/style"), and the latter would add 20 to the total for its assigned category ("personal task-based").

A further breakdown was made for the category "other language" to assess what languages had been used by users adding tags.

Finally, the type of book was taken into account in order to assess how categorisation varied across book type, for books targeted at young adults and for non-fiction books.

4.4.2 Overall categorisation results

Table 4-5 shows the breakdown of categories for all fifty analysed books, sorted by descending frequency on the "Tag Count & Freq. (%)" column. Both totals – the first taking

into account tag count ("Tag Count") and the second taking into account both tag count and frequency ("Tag Count & Freq."), are displayed. The "Tag Count (%)" column shows the percentage of the total of tag count for each category, and the "Tag Count & Freq. (%)" column shows the percentage of the total amount of "Tag Count & Freq." for each category.

Category	Tag Count	Tag Count (%)	Tag Count & Frequency	Tag Count & Frequency (%)
genre/style	1487	11.2%	158591	36.2%
Subject	4086	30.7%	120981	27.6%
personal task-based	2009	15.1%	32782	7.5%
target reader	675	5.1%	23653	5.4%
location	353	2.7%	17958	4.1%
author information	411	3.1%	16423	3.7%
Date	768	5.8%	15228	3.5%
physical item	690	5.2%	9317	2.1%
awards/popularity	320	2.4%	6940	1.6%
character/setting information	301	2.3%	6727	1.5%
opinion	520	3.9%	5599	1.3%
series information	112	0.8%	5266	1.2%
title information	78	0.6%	4370	1.0%
Code	547	4.1%	3189	0.7%
other language	357	2.7%	2942	0.7%
language of book	158	1.2%	2344	0.5%
movie information	97	0.7%	1817	0.4%
Blank	8	0.1%	1570	0.4%
publisher information	80	0.6%	973	0.2%
reading system	105	0.8%	843	0.2%
website	103	0.8%	514	0.1%
?	33	0.3%	191	0.0%
reference	12	0.1%	65	0.0%
translator/narrator/illustrator	11	0.1%	57	0.0%
Grand Total	**13321**	**100.00%**	**438340**	**100.0%**

Table 4-5 Categorisation summary for all tags

The highest number of tags was found to be in the genre/style category, at 36.1% of all tags. Subject also ranked highly, with 27.6% of all tags falling within this category. A sharp drop off occurs at this point, with the next most common type of tag being the "personal task-based" type, at 7.5% of the total, and target reader, location, author information and date all following with totals of between 3.5 and 5.4%. Categorisation based on tag count

Figure 4-2 shows the categorisation model with only a count of tags taken into account, with no account taken of the frequency of those tags.

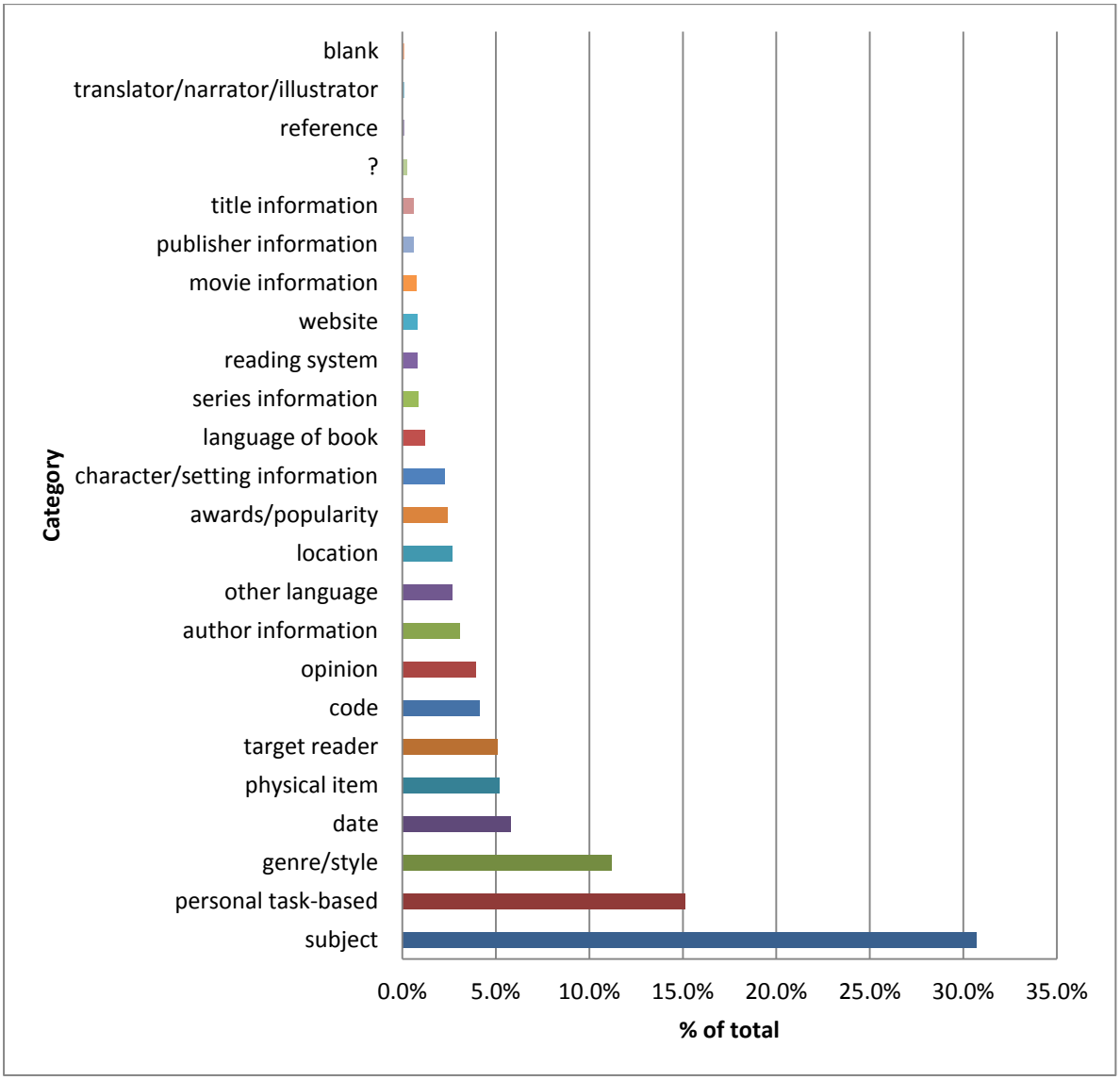

Figure 4-2 Categorisation based on tag count

4.4.3 Categorisation based on tag count combined with tag frequency

Figure 4-3 shows the categorisation model when both the count of individual tags but also the frequency of application of each tag was taken into account.

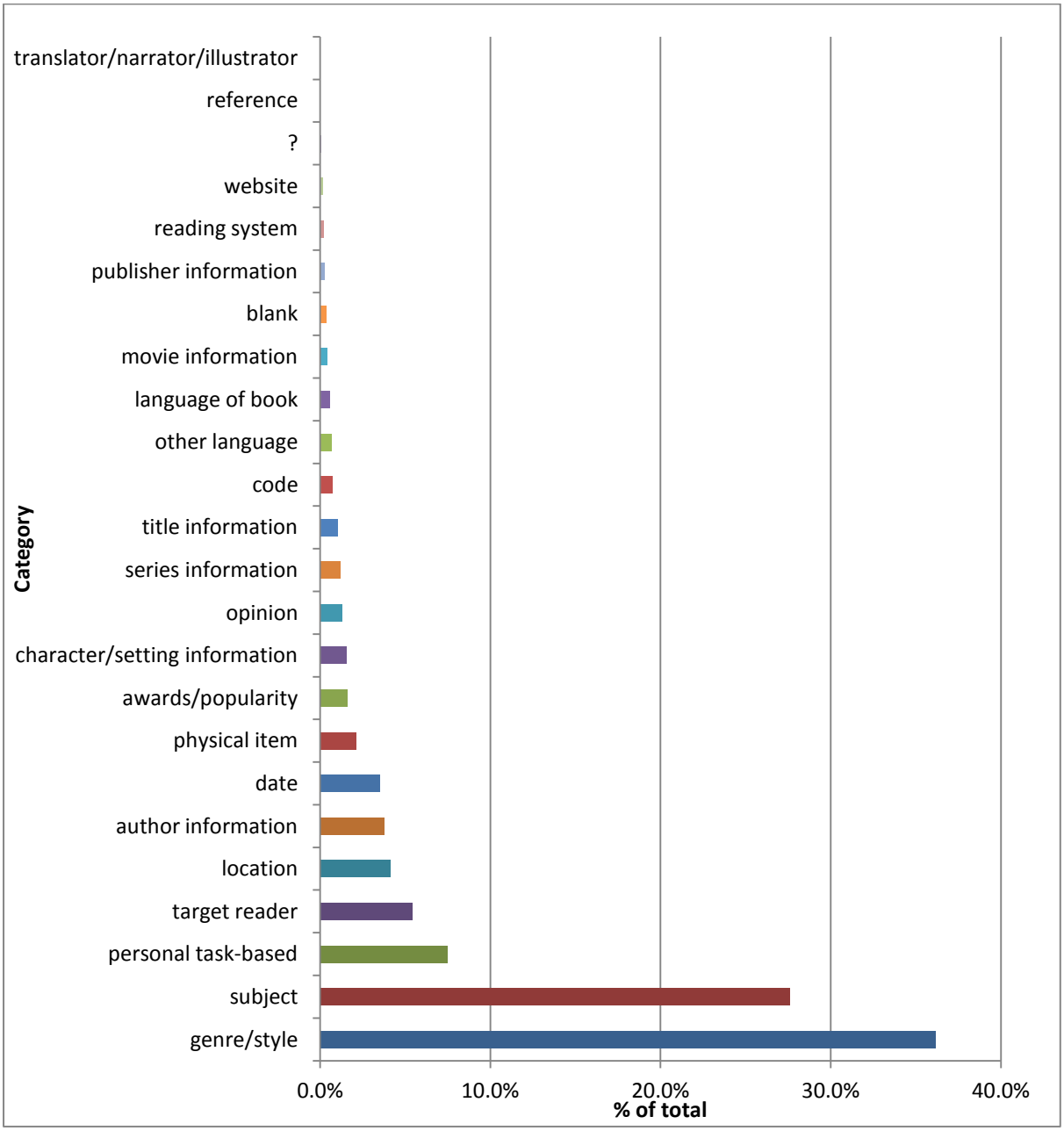

Figure 4-3 Categorisation based on tag count combined with tag frequency

4.4.4 The long tail of categorisation data

Ke & Chen (2012)'s concluded that the long tail of tag categories echoed a power law distribution. Plotting this study's categories against their frequencies without any scaling did indeed estimate a power law curve (see Chart 4-2), but once the scale was converted to a logarithmic scale in order to assess it for agreement with the Zipfian power law distribution, the distribution became linear rather than in power law form (see Chart 4-3). Ke & Chen's study did not use a logarithmic scale in order to generate the "power-law" curve, hence the use of the term "echoed" instead of "demonstrated". This means that this study does indeed replicate their study, but it would not be strictly true to say that the long tail is a formal Zipfian power law curve, as taking the log of the frequency alters the shape of the curve.

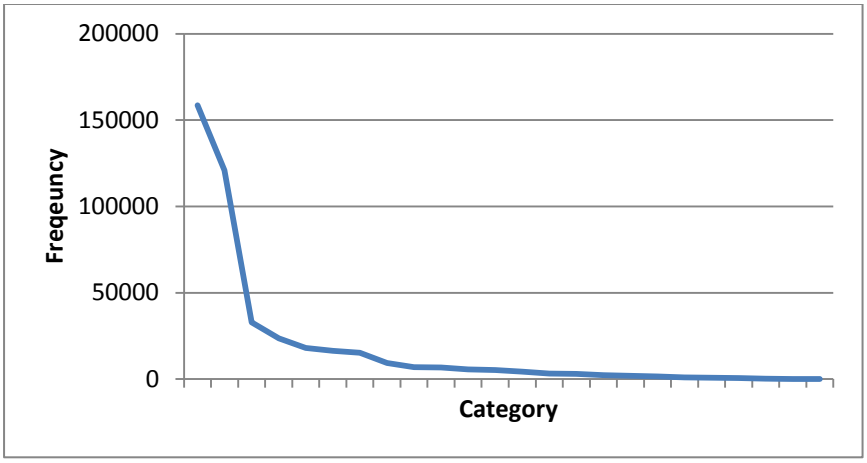

Chart 4-2 Categories plotted against category frequency

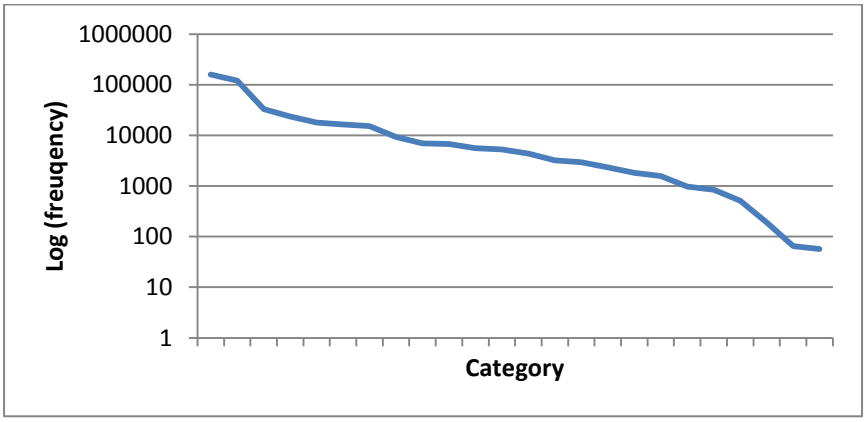

Chart 4-3 Categories plotted against log of category frequency

4.4.5 Categorisation according to book type

4.4.5.1 Young adult books

Of the fifty analysed books, nineteen are generally considered to be mainly aimed at the

young adult reader, although there is of course some considerable crossover into the adult market (see Appendix A: List of books for analysis).

Taking these books as a group, the categorisation changes somewhat:

Category	Tag Count	Tag Count & Frequency	Tag Count & Frequency (%)
genre/style	440	52707	32.2%
subject	1307	44832	27.4%
target reader	439	18956	11.6%
personal task-based	630	10602	6.5%
author information	111	5195	3.2%
character/setting information	199	4694	2.9%
series information	86	4638	2.8%
title information	36	3802	2.3%
physical item	244	3386	2.1%
date	194	3269	2.0%
location	74	2835	1.7%
opinion	197	1927	1.2%
awards/popularity	91	1441	0.9%
code	245	1404	0.9%
other language	131	990	0.6%
language of book	63	869	0.5%
movie information	44	815	0.5%
blank	2	486	0.3%
reading system	44	340	0.2%
publisher information	8	163	0.1%
Website	31	152	0.1%
?	19	130	0.1%
translator/narrator/illustrator	5	31	0.0%
reference	2	19	0.0%
Grand Total	4642	163683	100.0%

Table 4-6 Categorisation summary for tags on young adult books

4.4.5.2 Non-fiction books

Most of the fifty analysed books fall into the fiction genre, with six exceptions (see Appendix A: List of books for analysis). Again, the categorisation pattern changes when these books are analysed as a group:

Row Labels	Tag Count	Tag Count & Frequency	Tag Count & Frequency (%)
Subject	563	16721	47.0%
genre/style	111	10206	28.7%
personal task-based	190	3032	8.5%
location	39	1806	5.1%
date	75	983	2.8%
physical item	62	738	2.1%
opinion	47	715	2.0%
author information	31	529	1.5%
code	31	184	0.5%
blank		136	0.4%
target reader	15	124	0.3%
movie information	8	117	0.3%
other language	14	67	0.2%
awards/popularity	13	67	0.2%
title information	8	56	0.2%
character/setting information	5	36	0.1%
language of book	4	30	0.1%
website	5	22	0.1%
publisher information	2	9	0.0%
reading system	1	3	0.0%
Grand Total	1224	35581	100.0%

Table 4-7 Categorisation summary for tags on non-fiction books

4.5 Summary

In total, 13,358 tags were assessed and categorised during the study. These tags had been applied 438,340 times by various users. Tag data was found to follow a power law distribution, with the majority of tags consisting of only one word. A majority of tags, over

63%, consisted of only one word, with a further 27.7% being two word tags, and the most common tag length was seven characters, at just under 10% of all tags. Other language tags made up 0.7% of the dataset, with French being the most prevalent language in this set. The categorisation of the tags revealed that genre/style was the highest frequency category, at 36.2%, with subject following at 27.6%. Personal task-based tags, those tags most related to personal organisation and task completion, accounted for 7.5% of the final tags. As expected, taking frequency of application into account changed the landscape of the resulting category model a substantial amount when compared with a category model just based on tag count. Categorisation patterns were found to vary across book types with the examples taken being young adult and non-fiction books. Categorisation data took the form of a power law curve when linearly plotted, but not when the frequency of categories was converted to logarithmic form.

5 Discussion

5.1 Introduction

This chapter will discuss the results of the study with respect to the research objectives, namely the category model resulting from the analysis of the LibraryThing tags, the distributions of tags and categories with respect to frequency, and the tagging motivations the generated category model might suggest. How the study relates to the relevant literature will also be discussed, both in general, and particularly with respect to the category model and how it compares with other similar studies.

5.2 Tag Data

5.2.1 High frequency tags

Some individual tags were applied so frequently by users that they, on their own, represent a larger proportion of all tags than some of the categories in the category model themselves. Most of the highest frequency tags were categorised in the "genre/style" category. For example, the tag *fiction* was applied 54,190 times in total, on 49 books (surprisingly as 6 of the 50 sample books were considered non-fiction). This represented over 12% of total tag application. Other high-frequency tags included *fantasy* (5.9% of total) and *young adult* (2.2% of total). The tag *read* was the only one of the top ten high frequency tags that appeared on all books. *Young adult* was applied to 32 books, interestingly, despite the "official" young adult count of nineteen books. This, of course, probably reflects the fact that taggers know that it is not only books aimed specifically at them, that might be of interest or benefit to young adult readers.

5.2.2 The long tail of tag data

As predicted by Mathes (2004), the tag distribution in this study followed a typical Zipfian power law distribution. This confirms observations made in other studies, such as (Angus, Thelwall, & Stuart, 2008; Heymann & Garcia-Molina, 2009; Bischoff, Firan, Nejdl, & Paiu, 2008). What this indicates is a smaller group of popular tags accounting for a high proportion of all tagging activity, with a larger number of less popular tags that account for a low proportion each of tagging activity.

According to Guy & Tonkin (2006), "only ten to fifteen percent of the tags sampled on Flickr and del.icio.us are single-use tags". This statistic was borne out by the data within this research, where an average of 83% of the total number of tags (when tag frequency was

taken into account) were tags with a frequency of application of 3 or more, therefore only approximately 17% of the tags analysed were single or double-use tags.

Tagging data therefore falls into a distribution with two main components, the "short head" and the "long tail". In the short head lie the tags very likely to be used in tagging and also in search. As stated by Halpin et al (2007) "one can "safely ignore the "long-tail" of idiosyncratic and low frequency tags that are used by users to tweak their own results for personal benefit, or alternatively, treat the "long-tail" as an object of examination for other reasons" and Suchanek et al. (2008) "aggregating the top tags of a document biases to filtering out the meaningful tags". The more popular, high frequency tags are, in other words, the "bread and butter" of the tagging system, and are likely to be the most obvious tags to be placed on books (for example "genre/style" *fiction*, which on its own accounted for 12.4% of total tagging activity in the sample set).

However, while the popular tags are important tags, and will be highly used, the tags in the long tail do have their value, as noted by Shirky (2005) who commented that with such diversity of tags, a user need not wonder what the best search strategy to find a link might be, but instead ask "Is anyone tagging [the link] the way I do?".

If this study were to be continued, it would be interesting to carry out an analysis of the long tail frequency 1 and 2 tags, possibly carrying out a comparison with the category model generated from the analysis of the frequency 3 and higher tags.

5.2.3 Number of words per tag

The vast majority of tags (63%) consisted of only a single word, with two word tags accounting for a further 27.7% of tags. This correlates quite closely with the data from Heckner, Mühlbacher, & Wolff's 2008 study of tags on items in the Connotea database, in which approximately 71% of tags were single word tags, and 24% were two word tags.

5.2.4 Tag length

No other studies in the reviewed literature provided statistics for tag length. It might be of interest for future research to assess how single word tags compare with natural language in terms of length.

5.2.5 "Other language" tags

It is interesting to note that in terms of tag count, German is the most prevalent tagging language after English, but once tag frequency is taken into account, the French language tags

become more prominent. Essentially, there are more tags in German, but there are more people who apply the smaller number of French tags, than people who apply the larger number of German tags.

As to why these other languages appear on the English site at all, LibraryThing does offer localised websites (www.librarything.fr, www.librarything.de, www.librarything.nl, www.librarything.it) in four of the top five "other" languages represented (French, German, Dutch, Italian), with the only exception being Swedish, but one must assume that there is a group of users who prefer to use the English site, or these may possibly be users who joined LibraryThing before the localised site for their language was available.

5.2.6 Misspellings

Misspellings were not particularly prevalent in the sample set, with only a total of 0.02% of tags marked as misspelled. Intuitively, this makes sense - no tags with a frequency of 2 or lower were analysed, and although some misspellings are commonly duplicated (for example "recieve" instead of "receive"), most are likely to be non-duplicated spelling errors by a single user and thus not have a frequency higher than 1.

There have been varied results in the literature for the number of misspellings found in tags. For example, Thomas, Caudle, & Schmitz (2010) found that 5.24% of LibraryThing tags were misspelled and Guy & Tonkin (2006) found that 28% of del.ici.ious tags were in the category "misspellings, incorrect encodings and compound words". In both of these studies, however, the "long tail" of tags was taken into account, and the criteria for something being a misspelling were broader than in this study, with Guy & Tonkin's study, for example, including all non-English words in the "misspelling" category. Adding all non-English ("other language"), "?" and "code" tags to the misspelling total in this study gives a percentage of 1.42%. Taking this figure into account together with the fact that tags of frequency 1 and 2 were removed from the analysis would be likely to bring this study's misspelling result closer to these studies.

5.3 Categorisation Data

5.3.1 Overall categorisation results

It is clear from the categorisation results that taking tag frequency into account makes a substantial difference compared with only considering tag count. As the concluding chapter highlights in more detail, this aspect of the analytical models created for this study will

contribute to the ongoing academic debate, outlined in Chapter 2, about user motivation for tagging.

An example of this phenomenon is the "genre/style" category. While only representing 11.2% of all tags based on count, this category represents a very substantial 36.2% when frequency is also taken into account. Looking at the individual tags within this category, one can see that there are a large proportion of very high frequency tags within this set, such as *fiction* (appeared on 49 of the books, hence the tag count of 49, but with a total frequency of application of 158,591) and *fantasy* (appeared on only 28 of the books, hence the tag count of 28, but with a total frequency of application of 25,648), explaining the large jump in proportion of this category when the count is multiplied by the frequency. The "personal task-based" category halves in importance when tag frequency is taken into account, going from 15.1% to 7.5%. This can probably be explained by the fact that, as personal tags are more likely to be unique, there are more likely to be a higher number of individual tags, but they are less likely to be used by a substantial number of taggers (although each tag categorised was used by at least 3 taggers).

The "personal-task based" category is an interesting one, as it might be considered to apply only to the individual tagger themselves, but the patterns of high frequency of application show that individual taggers tend to follow collective patterns even in their personal tagging. The top ten tags from this category are shown in Table 5-1.

Tag (personal task-based)	Tag Count	Tag Count (%)	Tag Count & Frequency	Tag Count & Frequency (%)
read	50	0.4%	8439	1.9%
own	50	0.4%	3128	0.7%
TBR	46	0.4%	2490	0.6%
unread	48	0.4%	2404	0.5%
book club	39	0.3%	1333	0.3%
read in 2009	42	0.3%	701	0.2%
read in 2008	35	0.3%	695	0.2%
read in 2010	42	0.3%	657	0.1%
library	48	0.4%	645	0.1%
owned	36	0.3%	640	0.1%

Table 5-1Top ten "personal task-based" tags

5.3.2 Grouping of categories

It is useful to organise the categories into groups as indicated below in order to get a broader view of tagging behaviour and motivations. Percentages in parentheses beside group names indicate "Tag Count and Frequency (%)" figures for all categories assigned to that group.

Genre group (36.2%): Genre/style

Subject group (27.6%): Subject, character/setting information, reference

Metadata group (15.7%): Author information, date, location, awards/popularity, series information, title information, movie information, publisher information, translator/narrator/illustrator

Personal group (12.3%): Personal task-based, physical item, code, ?, language of book, opinion, website

Reader group (5.6%): Target reader, reading system

Miscellaneous group (1.1%): Other language, blank

The "date" and "location" categories could debatably belong in the personal group, rather than the metadata group, as some of the referenced dates and locations are likely to be based on personal information, rather than strict metadata information (for example, the tag *2008* for one tagger might be a reference to when a book was published, and for another tagger, might be a reference to when they purchased the book. This demonstrates the problem inherent in the categorisation process, namely a lack of context for tags. The "other language" category tags have been placed in the Miscellaneous grouping, rather than translating the tags and placing them into their appropriate categories. The "language of book" and "physical item" categories were included in the Personal grouping as the tags within these categories tend to refer to the actual book *object*, rather than the book as a general entity, and so seemed to fit better into the Personal group rather than the Metadata group.

5.3.3 The long tail of categorisation data

As shown in section 4.4.4, categorisation data did not fall into a tidy Zipfian power law distribution in log form, but did in non-log form. This is most likely to do with the relatively small number of categories, allowing the distribution to become apparent on the non-logarithmic scale. With a larger amount of data, the conversion to logarithmic data is necessary in order to facilitate the power law curve being visible on a reasonably sized graph. The non-logarithmic curve demonstrates the same underlying pattern as the logarithmic curve would have, that generally, the data falls into a pattern of the higher frequency categories having a much higher incidence than the lower frequency categories, with the curve tending to flatten out after the initial "short head" of very high frequency categories.

5.3.4 Comparison with categorisation models within the literature

5.3.4.1 Golder & Huberman (2006)

Although Golder & Huberman's study (2006) deals with a substantially different resource type (URLs), there are some similarities in the results. The "identifying what (or who) it is about" and "identifying qualities or characteristics" categories identified in that study have parallels with the "subject" and "opinion" categories in the current study respectively. The "personal task-based" category in the current study could be considered an amalgamation of Golder & Huberman's three categories "Identifying who owns it", "Self reference" and "Task organizing". As stated within their study "even information tagged for personal use can benefit other users". The "opinion" category contents are a good example of this, for example, the fact that multiple users have tagged a book as "favorite" is likely to make the tag a useful one for other users looking for a book recommendation.

5.3.4.2 Kipp (2007)

Kipp (2007) also found that personal task-based, or "non subject tags" made up a substantial proportion of the tags in a folksonomy. In her 2007 study, these types of tags made up 16% of all tags analysed. Interestingly, the "personal task-based" tags in this study were found to make up 15% of tags when only tag count was taken into account, but once tag frequency was included in the data, the percentage dropped dramatically to 7.5%. This makes intuitive sense, as duplicate personal tags are less likely to be applied by larger numbers of people than subject tags, as they tend to use personal terminology and references.

5.3.4.3 Heckner, Mühlbacher, & Wolff (2008)

Similarly to Kipp (2007), Heckner, Mühlbacher, & Wolff's study (2008) found that 20% of tags were "time and task related". They also found that within their study of internet resource tags, subject related tags could be further broken down into "resource related" and "content related" tags, which they calculated at percentages of 2% and 98% respectively. The "resource related" category has parallels with the "physical item" category within this study, which had a very similar percentage of 2.1%. Heckner et al also divided the "non-subject related tags" in their study into three sections – "affective", "time and task related" and "tag avoidance" – and found that these had percentages of 0.1%, 1.6% and 6.3% of the total respectively. The "time and task related" category within their study can be considered equivalent to the "personal task-based" category (7.5%) in this study, and the "affective" category to the "opinion" category (1.3%), so proportions in the two studies were quite

different. Again, it is difficult to pinpoint the exact reasons for this due to difference in tagging based on different types of resources, differences in methodology, subjectivity in categorisation, and so on.

5.3.4.4 Lawson (2009)

Lawson (2009) discovered that "subjective" (or non-content tags) generally fell into one of the following categories: "Reading Status", "Date", "Initials of tagger", "Type", "Gift suggestion", "Format", "Referral", "Location" (Lawson's study used a location category for tags such as "shelf in library" as opposed to this study's location category, which was reserved for geographical locations), "Bibliographic", "Opinion", "Author" and "Publisher". The "Format" tag correlates approximately with the "physical item" tag in this study, with the "Opinion", "Author", "Publisher" and "Date" tags also having obvious parallels with the "opinion", "author information", "publisher information" and "date" categories within the current study. Lawson's definition of "objective" tags corresponds to the "subject" category of the current study, which had a percentage of 27.7% as compared with Lawson's 20%.

5.3.4.5 Thomas, Caudle, & Schmitz (2010)

Within this study, 5.6% of tags analysed were found to be foreign language tags, 5.2% misspellings, and 5.6% dates. The figures for the current study are quite different, with 0.7% of tags being in a language other than English, only 0.02% misspelling tags, and 3.5% of tags falling in the "date" category. The comparatively low numbers for non-English and misspellings may be due to the fact that the long tail was not analysed in this study, but was in the Thomas et al study, and these types of tag are much more likely to have lower frequencies.

5.3.5 Categorisation according to book type

5.3.5.1 Young adult books

Of course, many adults read books that are aimed at the young adult market, but it can be assumed that more of the tags in the young adult book subset had been applied by young adult readers than those in the overall tag set.

Tags in the category "subject" had about the same importance in the set of tags for all books (27.6%), and in the subset for young adult books (27.4%). The genre/style category was represented in quite similar proportions also, with 36.2% overall, and 32.2% for young adult books. One tag category that showed quite a substantial difference was the "target

reader" category, with 5.4% of tags falling in this category generally, but 11.6% in the set of tags on young adult books. The percentage of "personal task-based" tags was also slightly higher (7.5%) in the full set of books than in the young adult set (6.5%). The "opinion" category had very similar proportions in both sets (1.3% versus 1.2%) – proving wrong an initial assumption that young adult taggers might have been more likely to use opinion tags. Surprisingly, exactly the same percentage of tags had the "reading system" category in both sets (0.2%) – it was expected that the young adult books might be more likely to be included in reading systems aimed at younger readers, and thus more likely to have associated tags applied.

5.3.5.2 Non-fiction books

As the non-fiction books only counted for six of the fifty books, it would not be statistically valid to draw conclusions from any small category differences that do occur. Most categories showed quite similar numbers for the two sets. However, a definite pattern seems to emerge in the "subject" category, accounting for 47% of tags in the non-fiction set but only 27.6% in the set overall. It could be inferred that the nature of the book itself impacts on the tags used and possibly even the motivation for tagging.

5.4 Motivations of taggers

As discussed in section 2.6, the main user motivations proposed by previous studies for tagging are organisation/categorisation and communication/description. According to Körner et al (2010), these two motivational types generally represent two distinct types of users (although users may be both categorisers and describers), the first who tend to use a smaller set of tags to succinctly describe resources and align them with an existing category model (the user's own or the folksonomy as a whole), and the second who tend to use a larger set of tags that more specifically describe resources. Description-type tags are often considered more useful for information retrieval due to a higher number of synonyms. According to the various studies outlined, most users tag mainly for personal purposes, but some also have a social purpose in mind as they apply tags.

It proved difficult to align the categories from this study's category model with the broad motivational groupings of organisation versus description, as the same tag could be considered to be an organisational tag or a description tag, depending on context. Some categories do align more closely with organisation (for example "target reader", "date", "publisher information") and others with description ("subject", "reference") but not to such

an extent that any definite conclusions about the validity of the organisation versus description model could be drawn.

Assessing personal versus social motivation is a more straightforward task. Broadly, it seems clear that users do have some social motivation in their tagging. For example, the fact that 1.3% of tags express an opinion would suggest that users believe that others might benefit from their tags. However, as the less social categories of "personal task-based" and "physical item", for example, account for a much larger percentage (9.6% combined), personal motivations would appear to be much more of a factor than social motivations. Of course, personal tagging can have social benefits, without the tagger necessarily intending them (for example if multiple taggers tag a book as "want to read", other users of the site who like similar books to those taggers, might use the tag to find a book recommendation).

5.5 Summary

The results of this study show that some tags have a very high frequency on their own, indeed one tag on its own (*fiction*) represented over 12% of all tag applications. The tag distribution was shown to demonstrate a Zipfian power law form, indicating that higher frequency tags in the "short head" of the curve have a much higher frequency than the large number of tags in the "long tail" of the curve. The decision not to analyse the "long tail" of tags in this study, in the case of this study meaning tags with a frequency of application of 1 or 2, was shown to be a sound decision by the statistic that 83% of tag applications were still analysed. However, the future possibility of analysing the long tail is discussed, due to the fact that useful tags and interesting patterns would be likely to emerge in the data, as shown in other similar studies that also included the long tail. Another future research possibility involves analysing tag length and how it compares with natural language data in general.

From the categorisation results, it is clear that taking frequency of application into account made a substantial difference within this study, compared with just assessing based on tag count. Grouping the twenty-four categories from the category model allows for a broader overview of the proportions of user tags, and indicates that proportions of user tags are as follows: genre group 36.2%, subject group 27.6%, metadata group 15.7%, personal group 12.3%, reader group 5.6%, and miscellaneous group 1.1%. The lack of context associated with tags was an issue when assigning categories to groups (as it was when assigning tags to categories). The categorisation distribution did not strictly follow a Zipfian power law curve, but did demonstrate the same features when charted linearly, namely that a

small number of categories were seen to have very high frequencies, and a larger number of categories were seen to have quite low frequencies. A comparison with the category models generated by the literature was carried out, and some parallels were drawn, but overall, it was found to be difficult to compare category models between studies effectively, due to the variability in data set size, resource type and categorisation rules. Young adult and non-fiction books were shown to show different patterns of categorisation than the book set overall, with young adult books, for example, demonstrating a much higher frequency of "target reader" tags than the set as a whole.

Finally, motivation of users was discussed, with the social and personal motivations the clearest to emerge, showing that users do show some social impulses behind their tagging, but mainly tag for their own purposes. The fact that personal tagging can have social benefits without the tagger necessarily intending them was also mentioned. Comparisons with the literature on organisation versus communication motivations were not conclusive, as it is very difficult to assess precisely which of these two motivations might be in question.

The difficulties in assessing the exact motivations for the use of particular tags, suggests that a qualitative component to the study would have been useful, with taggers asked about some of their tagging behaviours in order to gain a better insight into their motivations. Some examples of this type of qualitative analysis of tagging motivations were available in the relevant literature.

6 Conclusion

6.1 Introduction

This study set out to investigate and categorise the tags within a book tagging system, generating a category model that could help to assess the motivations of the users who applied the tags. The following chapter reviews the study, assesses whether the aims and objectives were met and discusses the study's findings. Limitations of the study are outlined, as are possibilities for future areas of research.

6.2 Aims and objectives

The study was carried out using a quantitative approach. The main aim of the research was to investigate the tags applied to books by website users in order to define a category model for a folksonomy specifically containing book tags (a "booksonomy"), with a view to understanding the motivations behind its users' tagging behaviours. The objectives which the research intended to address were:

- To undertake a review of the scholarly literature regarding folksonomies and user tagging decisions and purposes.
- To analyse a sample set of book tags applied by multiple users and to categorise those tags, thus building up a category model of a "booksonomy". The category model should take into account frequency of application of tags as well as tag counts.
- To assess whether both the tag and the category distributions follow the Zipfian power-law distribution model.
- To assess how the book tag category model compares with category models suggested by the literature, for books and for other resources.
- To discuss what the categorisation of the tags might imply about taggers' motivations when tagging books.

6.3 Literature review

A review of the literature in the field was carried out, with particular emphasis on the topics of folksonomies in general, how folksonomies compare with more formal ontological systems and how tags applied within tagging systems can be improved. Various categorisation studies across differing resource types were reviewed, with particular focus placed on studies specifying books as the resource type. Another important topic within the review of the literature was the assessment of user motivations for tagging.

Many of the studies within the literature did not take into account the frequency of application of a tag to a particular resource when categorising tags. This weighting of tags according to their popularity with users provides a more accurate category model, and thus the inclusion of the frequency of application of tags, together with the counts of tags, was an important addition to the research question, and thus the objectives of the study. The additional insight gained from combining both tag and frequency information within this study enhances the literature on tag distribution and categorisation.

6.4 Methodology

A quantitative approach was used within this study, with statistical analysis carried out on a large set of over 13,000 categorised tags in order to answer the components of the research question. Data was processed in various ways in order to clean and normalise it into a form that was conducive to robust and consistent categorisation. Once a detailed assessment had been made based on the literature and also based on the data itself, a decision was made to focus the analysis on the "short head" of the data, removing the "long tail" of tags with frequency of application of 1 and 2. This "short head" represented over 80% of total tag applications. An initial category model and categorisation process were decided upon based on a small sample set of books, and a revised category model and process were then used to carry out the full categorisation of the data. Microsoft Excel pivot tables were used to analyse the data.

6.5 Results and discussion

In total, 13,358 tags were assessed and categorised during the study, representing 438,340 tag applications. A power law distribution was observed in the tag data and approximated in the category data. Over 63% of tags were one-word tags, and the most common length of tag was seven characters. Non-English tags made up 0.7% of the total, with the top non-English language being French. Genre/style and subject were the two highest frequency categories, at 36.2% and 27.6% of total tag use respectively. 7.5% of all tags were "personal task-based" tags, such as *to read* and *finished in 2011*. Taking frequency of application into account or taking just tag counts into account, revealed substantially different category proportions, as expected. Book type also appears to have an effect on category proportions, with young adult books, for example, showing a higher proportion of "target reader" tags.

Some parallels with the category models from the literature were found, but it proved to be difficult to accurately compare categorisations between studies, due to the variances in

data sets, resource types and the rules for assigning tags to categories. Social and personal motivations of users who tag did emerge from the data, with the personal motivations generally appearing to substantially outweigh the social. The fact that personal tags can have social benefits even where the tagger does not have that direct intention, however, means that even personal tags contribute to the social value of the booksonomy. The assertions within the literature that tagging motivations usually break down into organisation versus communication, were also found to be difficult to assess within this study, as it was noted that many of the tags and tag categories could be associated with either motivation type, based on the definitions within the literature, and so a clear understanding of the proportions between them could not be ascertained.

6.6 Limitations

The quantitative nature of the study, while useful in its own right, does have its limitations. Without having context for a tag, it could be difficult to assign it to an appropriate category. Including a qualitative aspect to the study would have been helpful towards the categorisation itself and also towards understanding the motivations of users. Mixed-method studies within the literature could be considered to offer a more rounded overall view of tags and tagging motivations. Retrieving qualitative information for this data set would have been difficult, however, and so altering the approach might have overly limited the quantitative portion of the study.

Another limitation of the study is that as the categorisation was carried out by only one person, there was necessarily quite a high degree of subjectivity involved, with no opportunity for cross-referencing of categories applied. Furthermore, although the number of tags categorised was substantial, the fact that the study was only based on a set of fifty books, which themselves were chosen from a quite limited list of highly-reviewed books, means that the findings from the study cannot be assumed to extrapolate to all tags within all booksonomies.

6.7 Future research

As mentioned previously, carrying out this study using a mixed-method approach combining qualitative and quantitative data might allow for a more robust categorisation of tags and a more rounded understanding of taggers' motivations.

Another possibility for future research, given more availability of resources to analyse the set of tags, would be to carry out a full analysis including the long tail, to ascertain how that would affect category proportions and statistics such as number of misspellings.

Analysing how tag length and number of words per tag compare with other natural language data sets could also be an interesting future research approach.

A final topic that would be interesting for further research would be an analysis of how the type of book affects the category of tags that are applied to the book. Again, qualitative interview or survey information could be a useful addition to this type of study in order to ascertain if it is indeed the book type, or the demographics of the taggers that causes the difference.

6.8 Summary

Tagging is now a ubiquitous part of the information landscape, and tagging systems are required to be highly usable and robust. More and more information is coming on stream all the time, with fewer and fewer resources to annotate it, and so tagging systems must become ever more intelligent in order to use the valuable information provided by tagging users in order to build excellent search and recommendation systems. This study makes a contribution to the understanding of tagging within the field of books, and gives an insight into the contents of booksonomies and the motivations of people who tag books, all of which can assist designers in improving the usability and efficacy of book tagging, searching and recommendation systems.

7 Bibliography

Note: The sources in this bibliography are cited in APA (6th edition) format

Acronym Finder. (2012). Retrieved October-December 2012, from www.acronymfinder.com

Al-Khalifa, H. S., & Davis, H. C. (2007). Exploring the Value of Folksonomies for Creating Semantic Metadata. *International Journal on Semantic Web and Information Systems (IJSWIS), 3*, 13-39.

Al-Khalifa, H., & Davis, H. C. (2007). FAsTA: A Folksonomy-Based Automatic Metadata Generator. *Creating New Learning Experiences on a Global Scale*, 414-419.

Ames, M., & Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 971-980). New York: ACM.

Angus, E., Thelwall, M., & Stuart, D. (2008). General patterns of tag usage among university groups in Flickr. *Online Information Review, 32*(1), 89-101.

Bar-Ilan, J., Shoham, S., Idan, A., Miller, Y., & Shachak, A. (2006). Structured vs. unstructured tagging – A case study. *Collaborative Web Tagging Workshop at WWW2006.* Edinburgh, Scotland.

Bartley, P. (2009). Book tagging on LibraryThing: How, why, and what are in the tags? *Proceedings of the American Society for Information Science and Technology. 46(1),* pp. 1-22. Wiley Subscription Services, Inc., A Wiley Company.

Bates, J., & Rowley, J. (2011). 'Social reproduction and exclusion in subject indexing: A comparison of public library OPACs and LibraryThing folksonomy. *Journal of Documentation, 67*(3), 431-448.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008, 10 26). Can All Tags be Used for Search? *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pp. 193-202.

Bryman, A. (2004). *Social research methods* (2nd ed.). Oxford: Oxford University Press.

Cantador, I., Konstas, I., & Joemon, M. J. (2011). [ag_check] Categorising social tags to improve folksonomy-based recommendations. *Web Semantics: Science, Services and*

Agents on the World Wide Web, 9(1), 1-15. Retrieved from Web Semantics: Science, Services and Agents on the World Wide Web.

Garcia-Silva, A., Corcho, O., Alani, H., & Gómez-Pérez, A. (2012). Review of the state of the art: Discovering and Associating Semantics to Tags in Folksonomies. *The Knowledge Engineering Review, 27*(1), pp. 57-85.

Golbeck, J., Koepfler, J., & Emmerling, B. (2011). Experimental Study of Social Tagging Behavior and Image Content. *Journal of the American Society for Information Science and Technology, 62*(9), 1750-1760.

Golder, S. A., & Huberman, B. A. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science, 32*(2), 198-208.

Good, B. M., & Tennis, J. T. (2008). Evidence of Term-Structure Differences among Folksonomies and Controlled Indexing Languages. *Proceedings of the American Society for Information Science and Technology, 45*(1), pp. 1-7.

Google Inc. (2012). *Google Search*. Retrieved September-December 2012, from http://www.google.com

Google Inc. (2012). *Google Translate*. Retrieved October-December 2012, from Google Translate: http://translate.google.com/

Guy, M., & Tonkin, E. (2006). *Folksonomies: Tidying up Tags?* Retrieved September 1st, 2012, from D-Lib Magazine: http://www.dlib.org/dlib/january06/guy/01guy.html

Halpin, H., Robu, V., & Shepherd, H. (2007). The Complex Dynamics of Collaborative Tagging. *Proceedings of the 16th international conference on World Wide Web* (pp. 211-220). New York, NY, USA: ACM.

Heckner, M., Mühlbacher, S., & Wolff, C. (2008). Tagging tagging. Analysing user keywords in scientific bibliography management systems. *Journal of Digital Information*. Retrieved September 8th, 2012, from http://journals.tdl.org/jodi/index.php/jodi/article/view/246

Heckner, M., Neubauer, T., & Wolff, C. (2008). Tree, funny, to_read, google: what are tags supposed to achieve? A comparative analysis of user keywords for different digital

resource types. *Proceedings of the 2008 ACM Workshop on Search in Social Media* (pp. 3-10). New York, NY, USA: ACM.

Heymann, P., & Garcia-Molina, H. (2009). Contrasting Controlled Vocabulary and Tagging: Do Experts Choose the Right Names to Label the Wrong Things? *Second ACM International Conference on Web Search and Data Mining WSDM 2009, Late Breaking Results Session* (pp. 1-4). Stanford InfoLab.

Holley, R. (2010). 'Tagging Full Text Searchable Articles: An Overview of Social Tagging Activity in Historic Australian Newspapers August 2008 - August 2009. *D-Lib Magazine, 16*(1/2).

Irons, L. (2008). *Collective Tags, Collaborative Tags, the Long Tail, and Enterprise 2.0*. Retrieved September 8th, 2012, from SkilfulMinds [blog]: http://skilfulminds.com/2008/04/30/collective-tags-collaborative-tags-the-long-tail-and-enterprise-20/

Iyer, H., & Bungo, L. (2011). An examination of semantic relationships between professionally assigned metadata and user-generated tags for popular literature in complementary and alternative medicine. *Information Research, 16*(3).

Kakali, C., & Papatheodorou, C. (2010). Exploitation of folksonomies in subject analysis. *Library & Information Science Research, 32*(3), 192-202.

Kawase, R., & Herder, E. (2011). Classification of User Interest Patterns Using a Virtual Folksonomy. *Proceedings of 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL).* Ottawa, Canada.

Ke, H.-R., & Chen, Y.-N. (2012). Structure and pattern of social tags for keyword selection behaviors. *Scientometrics, 92*(1), 43-62.

Kipp, M. E. (2007). *@toread and Cool: Tagging for Time, Task and Emotion*. Retrieved September 8th, 2012, from http://dlist.sir.arizona.edu/1947/01/mkipp_iasummit2007.pdf

Kipp, M. E. (2010). User, Author and Professional Indexing in Context: An Exploration of Tagging Practices on CiteULike. *Canadian Journal of Information and Library Science, 35*(1), 17-48.

Kipp, M. E., & Campbell, D. G. (2006). Patterns and Inconsistencies in Collaborative Tagging Systems : An Examination of Tagging Practices. *In Annual General Meeting of the American Society for Information Science and Technology.* Austin, Texas.

Körner, C., Grahsl, H.-P., Kern, R., & Strohmaier, M. (2010). Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation. *21st ACM SIGWEB Conference on Hypertext and Hypermedia* . Toronto, Ontario, Canada: ACM.

Laniado, D., Eynard, D., & Colombetti, M. (2007). Using WordNet to turn a folksonomy into a hierarchy of concepts. *Semantic Web Application and Perspectives - Fourth Italian Semantic Web Workshop*, (pp. 192-201).

Lawson, K. G. (2009). Mining Social Tagging Data for Enhanced Subject Access for Readers and Researchers. *The Journal of Academic Librarianship, 35*(6), pp. 574-582.

LibraryThing. (2012). *LibraryThing*. Retrieved September 8-15, 2012, from LibraryThing: www.librarything.com

Lim, W. H., Alhashmi, S. M., & Siew, E.-G. (2011). The Information Potential and Temporal Elements of Web 2.0 Folksonomy for User Profiling in Personalized Information Retrieval. *Journal of Internet Social Networking & Virtual Communities*.

Lipczak, M. (2008). Tag recommendation for folksonomies oriented towards individual users. *Proceedings of the ECML PKDD Discovery Challenge*.

Lu, C., Park, J.-r., & Hu, X. (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science, 36*(6), 763-779.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (pp. 31-40). ACM Press.

Mathes, A. (2004). *Folksonomies - Cooperative Classiciation and Communication Through Shared Metadata*. Retrieved August 28, 2012, from www.adammathes.com: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Merholz, P. (2004). *Metadata for the Masses.* Retrieved August 20, 2012, from www.adaptivepath.com: http://www.adaptivepath.com/ideas/e000361

Morrison, P. J. (2007, October/November). Why Are They Tagging, and Why Do We Want Them To? *Bulletin of the American Society for Information Science and Technology*, pp. 12-15.

Overell, S., Sigurbjörnsson, B., & van Zwol, R. (2009). Classifying Tags using Open Content Resources. *WSDM '09 Proceedings of the Second ACM International Conference on Web Search and Data Mining*, (pp. 64-73). Barcelona, Spain.

Peters, I., & Weller, K. (2008). Tag Gardening for Folksonomy Enrichment and Maintenance. *Webology, 5*(3). Retrieved from Webology.

Pickard, A. (2007). *Research methods in information.* London: Facet Publishing.

Porter, J. (2005). *Controlled Vocabularies Cut Off the Long Tail*. Retrieved September 20, 2012, from Bokardo.com: Social Web design [blog]: http://bokardo.com/archives/controlled_vocabularies_long_tail/

Rader, E., & Wash, R. (2008). Influences on tag choices in del.ici.ious. *Proceedings of the ACM 2008 conference on Computer Supported Cooperative Work* (pp. 239-248). New York: ACM.

Ransom, N., & Rafferty, P. (2011). Facets of user-assigned tags and their effectiveness in image retrieval. *Journal of Documentation, 67*(6), pp. 1038-1066.

Shiri, A. (2007). Trend Analysis in Social Tagging : An LIS Perspective. *The 6th European Networked Knowledge Organization Systems (NKOS) Workshop at the 11th European Conference on Research and Advanced Technology for Digital Libraries.* Budapest, Hungary.

Shirky, C. (2005). *Ontology is Overrated: Categories, Links, and Tags*. Retrieved September 8, 2012, from Clay Shirky's Writings About the Internet: Economics & Culture, Media & Community [blog]: http://www.shirky.com/writings/ontology_overrated.html

Smith, G. (2008, August/September). Tagging: Emerging Trends. *Bulletin of the American Society for Information Science and Technology, 34*(6).

Spiteri, L. F. (2007). The structure and form of folksonomy tags: the road to the public library catalogue. *Information Technology and Libraries, 26*(3), pp. 13-25.

Suchanek, F. M., Vojnovic, M., & Gunawardena, D. (2008). Social Tags: Meaning and Suggestions. *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 223-232). New York, NY, USA: ACM.

Thomas, M., Caudle, D. M., & Schmitz, C. (2010). Trashy tags: problematic tags in LibraryThing. *New Library World, 111*(5/6), pp. 223-235.

Tonkin, E. (2006). Searching the long tail: Hidden structure in social tagging. (J. Furner, & J. T. Tennis, Eds.) *Advances in Classification Research, Volume 17; Proceedings of the 17th ASIS&T Classification Research Workshop*. Retrieved September 20th, 2012, from http://dlist.sir.arizona.edu/1791/01/tonkin.pdf

Tourné, N., & Godoy, D. (2012). Evaluating tag filtering techniques for web resource classification in folksonomies. *Expert Systems with Applications, 39*(10), 9723-9729.

Trant, J. (2009). Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information, 10*(1).

Vander Wal, T. (2007, February 2). *Folksonomy*. Retrieved August 28, 2012, from www.vanderwal.net: http://vanderwal.net/folksonomy.html

Weinberger, K., Slaney, M., & van Zwol, R. (2008). Resolving Tag Ambiguity. *Proceedings of the 16th ACM international conference on Multimedia* (pp. 111-120). Vancouver, British Columbia, Canada: ACM.

Wichowski, A. (2009). 'Survival of the Fittest Tag: Folksonomies, Findability, and the Evolution of Information Organization. *First Monday*, *14*(5).

Wikipedia. (2012). *Wikipedia*. Retrieved September-December 2012, from http://en.wikipedia.org

Zipf, G. K. (1935). *The Psychobiology of Language.* Houghton-Mifflin.

Zubiaga, A., Martínez, R., & Fresno, V. (2012). Analyzing Tag Distributions in Folksonomies for Resource Classification. *CoRR*.

Appendix A: List of books for analysis

	Title	Author	YA/NF?
1	The Book Thief	Markus Zusak	YA
2	The Da Vinci Code	Dan Brown	
3	New Moon	Stephenie Meyer	YA
4	The Lovely Bones	Alice Sebold	YA
5	The Lightning Thief	Rick Riordan	YA
6	The Thirteenth Tale	Diane Setterfield	
7	Life of Pi	Yann Martel	
8	The Boy in the Striped Pyjamas	John Boyne	YA
9	The Lost Symbol	Dan Brown	
10	Jane Eyre	Charlotte Brontë	
11	Harry Potter and the Order of the Phoenix	J. K. Rowling	YA
12	The Hobbit	J. R. R. Tolkien	YA
13	Sarah's Key	Tatiana de Rosnay	
14	Holes	Louis Sachar	YA
15	Three Cups of Tea: One Man's Mission to Promote Peace ... One School at a Time	Greg Mortenson	NF
16	Diary of a Wimpy Kid	Jeff Kinney	YA
17	The Memory Keeper's Daughter	Kim Edwards	
18	The Eyre Affair	Jasper Fforde	
19	Good Omens: The Nice and Accurate Prophecies of Agnes Nutter, Witch	Neil Gaiman	
20	The Art of Racing in the Rain: A Novel	Garth Stein	
21	Matched	Ally Condie	YA
22	The Poisonwood Bible	Barbara Kingsolver	
23	Pride and Prejudice and Zombies	Jane Austen	
24	The Immortal Life of Henrietta Lacks	Rebecca Skloot	NF
25	Snow Flower and the Secret Fan	Lisa See	
26	Marley & Me	John Grogan	NF
27	Anna Karenina	Leo Tolstoy	
28	Night	Elie Wiesel	
29	Miss Peregrine's Home for Peculiar Children	Ransom Riggs	
30	Hush, Hush	Becca Fitzpatrick	YA
31	The White Tiger	Aravind Adiga	
32	Lamb : The Gospel According to Biff, Christ's Childhood Pal	Christopher Moore	
33	Emma	Jane Austen	
34	The God Delusion	Richard Dawkins	NF
35	Hatchet	Gary Paulsen	YA
36	The Physick Book of Deliverance Dane	Katherine Howe	
37	A Confederacy of Dunces	John Kennedy Toole	
38	The Forgotten Garden	Kate Morton	

39	Bridge to Terabithia	Katherine Paterson	YA
40	The Name of the Rose	Umberto Eco	
41	Soulless	Gail Carriger	
42	Sworn to Silence	Linda Castillo	
43	Because of Winn-Dixie	Kate DiCamillo	YA
44	Switched	Amanda Hocking	YA
45	The Sea of Monsters	Rick Riordan	YA
46	Running With Scissors	Augusten Burroughs	NF
47	Evermore	Alyson Noël	YA
48	Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	Mitch Albom	NF
49	Stargirl	Jerry Spinelli	YA
50	Before I Fall	Lauren Oliver	YA

Appendix B: Statistics for the set of analysed books

Title	Count of all tags	Sum of frequency of all tags	Count of non-long tail tags	Sum of frequency of non-long tail tags	% of tags that were not in the long tail (count)	% of tags that were not in the long tail (sum of frequency)
The Book Thief	3247	19375	414	16268	12.8%	84.0%
The Da Vinci Code	4350	29431	594	25291	13.7%	85.9%
New Moon	3041	20620	432	17733	14.2%	86.0%
The Lovely Bones	2629	13960	370	11473	14.1%	82.2%
The Lightning Thief	2403	11699	339	9427	14.1%	80.6%
The Thirteenth Tale	2038	9825	273	7889	13.4%	80.3%
Life of Pi	3343	18966	466	15782	13.9%	83.2%
The Boy in the Striped Pyjamas	1503	6362	182	4928	12.1%	77.5%
The Lost Symbol	1381	6624	194	5342	14.0%	80.6%
Jane Eyre	4384	28153	592	23923	13.5%	85.0%
Harry Potter and the Order of the Phoenix	4935	43986	732	39294	14.8%	89.3%
The Hobbit	4999	40571	762	35760	15.2%	88.1%
Sarah's Key	1212	5069	169	3943	13.9%	77.8%
Holes	2741	10996	401	8378	14.6%	76.2%
Three Cups of Tea: One Man's Mission to Promote Peace ... One School at a Time	1903	8391	221	6547	11.6%	78.0%
Diary of a Wimpy Kid	1589	5601	215	4082	13.5%	72.9%
The Memory Keeper's Daughter	1736	7398	244	5761	14.1%	77.9%
The Eyre Affair	2098	12947	313	10977	14.9%	84.8%
Good Omens: The Nice and Accurate Prophecies of Agnes Nutter, Witch	2274	17974	353	15821	15.5%	88.0%
The Art of Racing in the Rain: A Novel	944	3784	137	2905	14.5%	76.8%
Matched	732	3013	110	2342	15.0%	77.7%
The Poisonwood Bible	2387	12604	335	10365	14.0%	82.2%
Pride and Prejudice and Zombies	1114	5891	166	4836	14.9%	82.1%

The Immortal Life of Henrietta Lacks	1339	6496	184	5237	13.7%	80.6%
Snow Flower and the Secret Fan	1505	6235	190	4798	12.6%	77.0%
Marley & Me	1322	6093	174	4836	13.2%	79.4%
Anna Karenina	3014	18192	391	15265	13.0%	83.9%
Night	2052	11882	268	9891	13.1%	83.2%
Miss Peregrine's Home for Peculiar Children	770	3403	106	2668	13.8%	78.4%
Hush, Hush	535	2083	68	1572	12.7%	75.5%
The White Tiger	1230	5081	174	3934	14.1%	77.4%
Lamb : The Gospel According to Biff, Christ's Childhood Pal	1293	6045	191	4832	14.8%	79.9%
Emma	2890	18719	380	15922	13.1%	85.1%
The God Delusion	1683	9616	223	8002	13.3%	83.2%
Hatchet	1810	6952	259	5251	14.3%	75.5%
The Physick Book of Deliverance Dane	771	3121	117	2395	15.2%	76.7%
A Confederacy of Dunces	2005	9462	270	7553	13.5%	79.8%
The Forgotten Garden	923	3050	126	2187	13.7%	71.7%
Bridge to Terabithia	2114	9466	314	7418	14.9%	78.4%
The Name of the Rose	2994	17798	426	14960	14.2%	84.1%
Soulless	775	4570	115	3839	14.8%	84.0%
Sworn to Silence	343	1123	48	790	14.0%	70.3%
Because of Winn-Dixie	1750	6170	267	4538	15.3%	73.5%
Switched	196	813	43	646	21.9%	79.5%
The Sea of Monsters	1447	6871	214	5497	14.8%	80.0%
Running With Scissors	1344	7013	200	5742	14.9%	81.9%
Evermore	561	2157	89	1618	15.9%	75.0%
Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson	1693	6838	227	5217	13.4%	76.3%
Stargirl	1269	4569	200	3387	15.8%	74.1%
Before I Fall	523	1768	80	1278	15.3%	72.3%
Total	**95134**	**528826**	**13358**	**438340**	**14.0%**	**82.9%**

Appendix C: List of possible categories

Possible categories
?
author information
awards/popularity
blank
character/setting information
code
date
genre/style
language of book
location
movie information
opinion
other language
personal task-based
physical item
publisher information
reading system
reference
series information
subject
target reader
title information
translator/narrator/illustrator
website

Appendix D: Initial categorisation example

Initial categories for “The Lovely Bones” by Alice Sebold:

Category	Sum of Frequency	Number of Tags	% of Total
subject	4403	102	38.4%
genre	3869	52	33.7%
action	791	32	6.9%
physical item	437	30	3.8%
date	301	14	2.6%
opinion	237	25	2.1%
action & date	183	11	1.6%
target reader	124	3	1.1%
reason for reading	115	7	1.0%
author information	111	11	1.0%
location/author information	106	1	0.9%
location	89	5	0.8%
location & genre	89	2	0.8%
subject/target reader	86	4	0.7%
movie information	85	5	0.7%
code	76	21	0.7%
award	63	11	0.5%
blank	54	1	0.5%
?	44	9	0.4%
genre & target reader	40	3	0.3%
subject & genre	31	3	0.3%
title	22	4	0.2%
character information	21	4	0.2%
date set	20	1	0.2%
other language - french	20	1	0.2%
popularity	19	2	0.2%
language of book	17	3	0.1%
other language - german	8	1	0.1%
other language - swedish	6	1	0.1%
other language - spanish	6	1	0.1%
Grand Total	**11473**	**370**	**100.0%**

Refined categories for the same book:

Category	Sum of Frequency	Number of Tags	% of Total
subject	4412	104	38.5%
genre/style	4210	66	36.7%
personal task-based	1314	65	11.5%
date	321	15	2.8%
opinion	237	25	2.1%
physical item	208	14	1.8%
location	195	6	1.7%
author information	111	11	1.0%
awards/popularity	86	14	0.7%
movie information	85	5	0.7%
code	76	21	0.7%
?	61	12	0.5%
blank	54	1	0.5%
other language	40	4	0.3%
target reader	35	2	0.3%
title information	22	4	0.2%
character information	6	1	0.1%
Grand Total	**11473**	**370**	**100.0%**